N° d'ordre : 3560

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Emmanuelle BEYNE**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Règles de cohérence pour l'annotation génomique : développement et mise en œuvre *in silico* et *in vivo*

Soutenue le : 17 Janvier 2008

Après avis des rapporteurs :

Médigue Claudine	Directrice de Recherche CNRS - Génoscope
Feldmann Horst	Professeur Émerite - Université de Munich

Devant la commission d'examen formée de :

Vauquelin Bernard ...	Professeur - Université de Bordeaux 1	Président de jury
Médigue Claudine	Directrice de Recherche CNRS - Génoscope	Rapporteur
Feldmann Horst	Professeur Émerite - Université de Munich	Rapporteur
Gaillardin Claude	Professeur - INRA Paris Grignon	Examinateur
Mosbah Mohamed ...	Professeur - ENSEIRB	Examinateur
Sherman David James	Maître de Conférences - ENSEIRB	Directeur de thèse

– 2008 –

Remerciements

Je tiens à remercier les différentes personnes qui m'ont permis de réaliser cette thèse ou que j'ai simplement croisées. Je ne citerai pas tout le monde car il y en a trop...

Dans la famille "officielle", je remercie en premier lieu David Sherman, mon directeur de thèse, et Pascal Durrens, mon conseiller bio (qui dit bio dit moustaches), de m'avoir proposé cette véritable aventure, aussi bien professionnelle qu'humaine.

Dans la famille "humide", je remercie Marc Bonneu (donc à moustaches) de m'avoir accueillie dans son laboratoire de protéomique, me permettant de retourner à la paillasse. Merci à toute la jeune équipe "humide" de m'avoir aidée à réaliser les manips, de leurs conseils, leur secours : Delphine, miss Labo qui m'a tirée de nombreuses situations périlleuses, Jean-Paul, mon encadrant bio, Stéphane et Émilie pour la partie spectro.

Dans la famille "mi-sèche, mi-humide", je remercie Antoine de Daruvar, directeur du CBiB de m'avoir hébergée au temps des manips. Merci aux personnes que j'ai croisées là-bas de leurs discussions, leurs encouragements, mais aussi les bons moments, les repas du midi (parfois) très animés, les secrets bien gardés : Monique, Hélène, Claire, Nico, Alexandre, Naser, Aurélien...

Dans la famille "sèche", je remercie l'ensemble des (ex-)membres bio-info du LaBRI de leurs conseils, leur aide et leurs encouragements (surtout à la fin !) : Tiphaine, Florian, Hayssam, Cyril, Géraldine, Isabelle L., Isabelle D., Aïda, Nicolas L., Pascal F., Macha, Roland... Avec tous mes encouragements pour les suivant(e)s...

Merci également aux diverses personnes que j'ai croisées au LaBRI, dans l'atrium, autour d'un verre, sur un parquet, dans les airs, et qui m'ont fait oublier la dure vie d'un thésard : Sylvie, Yvan, David R., Pierre R., Nicolas B., Afif, Aurel, Pierre H., Yon... Merci aux membres du GT Bio-info appliqué...

Dans la famille "levure", je remercie l'ensemble des personnes du consortium Génolevures, de leurs discussions, leurs conseils pour ce travail. Merci de m'avoir fait confiance dès mon stage de DESS.

Dans la famille "de sang", je remercie de tout cœur mes parents qui m'ont permis, ainsi qu'à mes sœurs, de faire des études, de nous avoir encouragées dans nos choix. Merci Raphaële et Marion, mes deux *petites* sœurs, de votre soutien. Merci à mes taties Confiture (Lili, Janine et Suzy) de les soutiens oraux, epistolaires et culinaires.... Merci en général à ma grande famille de leur soutien : oncles, tantes, cousins, cousines (une partie de ma famille est d'origine italienne alors ça fait du monde)...

Dans la famille "de cœur", je remercie de tout mon cœur mon Lord Bertrand. Merci pour tout ce que tu m'apportes, et tout le reste. Pour ton écoute, tes conseils, tes petits plats, tes mails (à defaut de lettre), tes encouragements... Merci aux parentoux, frérot, mamies Cirou et Viciana pour ce qu'ils m'ont apporté.

Et dans la famille "essentielle à la survie d'un thésard", je remercie Dame Nature d'avoir inventé notamment le cacaoyer et la levure, ainsi que leurs premiers consommateurs bipèdes...

à mes parents

Résumé

L'annotation génomique identifie l'ensemble des éléments significatifs présents sur l'ADN génomique, le support du programme de fonctionnement de l'organisme. Elle prédit leurs fonctions biologiques et leurs relations. L'annotation d'un génome complet est soumise à diverses contraintes : elle doit être réalisée rapidement et représenter l'organisme comme système biologique fonctionnel cohérent.

Nous proposons une méthode de vérification de la qualité de l'annotation génomique, basée sur un ensemble de règles de cohérence définies d'après les connaissances et contraintes biologiques admises par la communauté scientifique. Ces règles vérifient la complétude de l'annotation (présence des éléments vitaux pour l'organisme) et son absence d'erreur (sens biologique correct des éléments décrits).

Notre méthode est appliquée dans le cadre du projet Génolevures, un projet de génomique comparée chez les levures hémiascomycètes. Nous avons mis en place un système d'annotation facilitant le travail d'annotation manuelle par les experts. L'intégration de nos règles dans ce système permet de garantir la bonne qualité de l'annotation produite.

Nous avons choisi de valider expérimentalement l'application de ces règles en étudiant les interactions protéine-protéine chez les levures *Saccharomyces cerevisiae* et *Yarrowia lipolytica* par la technique de l'électrophorèse en gel de polyacrylamide en bleu natif et SDS (BN/SDS PAGE). Les résultats obtenus apportent de nouvelles connaissances chez les levures étudiées. Ils démontrent l'universalité de certaines règles et le bien fondé de la stratégie d'annotation.

Mots-clefs : annotation, génome, règles de cohérence, bio-informatique, interaction protéine-protéine, électrophorèse BN/SDS

Discipline : Informatique

LaBRI,
Université Bordeaux 1,
351, cours de la libération
33405 Talence Cedex (FRANCE)

Abstract

Genome annotation identifies all significant elements located on genomic DNA, support on which the functional program for organism is written. It predicts their biological functions and their relationships. The annotation of a completely sequenced genome is subject to several constraints : it must be done quickly and represent the organism as a coherent biological functionnal system.

We propose a method to check the quality of genome annotation, based on a set of coherency rules defined according to knowledge and constraints accepted by the scientific community. These rules check, on the one hand, the completeness of the annotation (presence of vital functions), and on the other hand, the absence of errors (correctness of biological sense of predicted elements).

Our method is applied in the Génolevures project, a comparative genomics project within the hemiascomycetous yeasts. We developed the annotation system that eases the manual annotation work by experts. The integration of our rules within this system garantees the good quality of the resulting annotation.

We further chose to validate experimentaly these rules with studying protein-protein interactions of yeasts *Saccharomyces cerevisiae* and *Yarrowia lipolytica*, using blue native and SDS polyacrylamide gel electrophoresis (BN/SDS PAGE). The results offer both new biological knowledge for the studied yeasts and a demonstration of the universality of certain rules as well as the well-founded of the annotation strategy.

Keywords : annotation, genome, coherency rules, bioinformatics, biocomputing, protein-protein interaction, BN/SDS PAGE

Discipline : Computer Science

LaBRI,
Université Bordeaux 1,
351, cours de la libération
33405 Talence Cedex (FRANCE)

Liste des symboles et abréviations

Symboles

§ : paragraphe
$\wedge$: et logique
$\vee$: ou logique
$\exists$: il existe
$\forall$: pour tout
$\cup$: union
$\cap$: intersection
$\in$: appartient à
$\times$ g : accélération, $1 \times g = 9.806\ m.s^{-2}$

Molécules chimiques

CH_3COOH : acide acétique
H_2O : eau
HCOOH : acide formique ou méthanoïque
NaH_2PO_4 : phosphate de sodium monobasique
NH_4HCO_3 : monocarbonate d'ammonium

Abréviations

aa : Acide Aminé
ABC : Cassette de liaison à l'ATP (ATP Binding Cassette)
ACN : Acétonitrile
ADN : Acide DésoxyriboNucléique
ADNc : Acide DésoxyriboNucléique complémentaire
ARN : Acide RiboNucléique
ARNm : Acide RiboNucléique messager
ARNpré-m : Acide RiboNucléique pré-messager
ARNt : Acide RiboNucléique de transfert
ATP : Adénosine TriPhosphate
BLAST : Outil de recherche d'alignement local de base (Basic Local Alignment Search Tool)
BN : Bleu Natif (Blue Native)
CDS : Séquence d'ADN codante (Coding DNA Sequence)
Da : Dalton, unité de masse protéique
di : Diamètre interne
DTA : Format de fichiers de données de valeurs numériques
DTT : Dithiothreitol
EST : Marqueur de séquence exprimée (Expressed Sequence Tag)
GO : Gene Ontology

HPLC : Chromatographie en phase liquide à haute performance (High-Performance Liquid Chromatography)

IPP : Interaction Protéine-Protéine

kb : kilo bases (kbp : kilo paires de bases)

LC-MS/MS : Chromatographie en phase liquide couplée à la spectrométrie de masse en tandem (Liquid Chromatography, Spectrometry Mass)

m/z : Rapport masse sur charge

nt : Nucléotide

ORF : Cadre ouvert de lecture (Open Reading Frame)

PAGE : Électrophorèse en gel de polyacrylamide (PolyAcrylamide Gel Electrophoresis)

pb : Paire de bases (1 Mpb = 10^6 pb)

PMSF : Phényl Méthyl Sulfonyl Fluoride

PTM : Modification post-traductionnelle (Post-Translational Modification)

qsp : Quantité suffisante pour

rpm : Rotation par minute

SDS : Sel de sodium de l'acide dodécyl sulfonique (Sodium Dodecyl Sulfate)

TEMED : TÉtraMéthyl Éthylène Diamine

UTR : Région non traduite (UnTranslated Region)

w/v : Unité de masse/ unité de volume = kg/l (weight/volume)

Xcorr : Score de corrélation m/z

YPD : Milieu de culture complet pour la croissance de levures (Yeast Peptone Dextrose)

YP_2DH_5 : Milieu de culture complet pour la croissance de levures enrichi en lipide

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

La découverte de la molécule d'ADN [Watson and Crick, 1953b, Watson and Crick, 1953a] a été fondamentale pour la biologie moderne car cette molécule est le support du programme de fonctionnement de la cellule. La compréhension précise du fonctionnement des micro-organismes vivants a d'importantes implications et répercussions dans les domaines tels que la médecine (amélioration de l'hygiène et de la santé, traitement des maladies ...), l'alimentation (amélioration de la qualité et du rendement), l'environnement ou les conditions de vie. Décrypter la molécule d'ADN d'un organisme et son rôle dans la machinerie cellulaire suscite ainsi l'intérêt de la communauté scientifique, mais aussi des industriels (agro-alimentaires, pharmaceutiques ...) et des instances politiques.

Connaître la composition de cette molécule ne suffit pas pour autant à la rendre informative. Pour avoir un sens, elle doit être analysée, annotée. L'*annotation génomique* est l'identification de l'ensemble des éléments significatifs du matériel génétique et de leur rôle. Elle fournit les informations nécessaires pour la compréhension du fonctionnement de la cellule et des relations entre les gènes, et se décline en trois niveaux :

- syntaxique, qui identifie les régions de séquences significatives sur l'ADN,
- fonctionnel, qui associe une fonction biologique à chacun de ces éléments et
- relationnel, qui établit les relations entre les éléments.

Historiquement, l'identification portait uniquement sur l'ensemble des gènes (ou génome) d'un organisme car les projets d'annotation ciblaient les protéines, produits d'expression des gènes. L'annotation relationnelle est apparue il y a une dizaine d'années [Goffeau et al., 1996, Venter et al., 2001] lorsque les premiers génomes entièrement séquencés disposaient de données en quantité suffisante pour établir des relations (*e.g.* les voies métaboliques) entre les éléments du génome. Désormais, l'annotation génomique intègre également l'identification de toute séquence ADN ayant une signification biologique.

Grâce au développement de la bio-informatique, l'annotateur dispose d'outils d'aide à

l'annotation et de méthodes d'analyses de résultats expérimentaux et prédictifs. La bio-informatique est un domaine de recherche fondé sur les concepts et les formalismes issus de la biologie, l'informatique, les mathématiques, la physique et la chimie. Cette discipline traite de l'analyse de l'information biologique, essentiellement sous la forme de séquences nucléiques, de séquences d'acides aminés et de structures de protéines. Par abus de langage, le terme bio-informatique fait également référence aux applications informatiques développées lors de ces recherches[1].

Pour une espèce donnée, l'annotation génomique peut être améliorée à chaque niveau en prenant en compte les connaissances d'espèces apparentées. Ainsi, la comparaison de génomes apparentés annotés permet de mettre en évidence les différences fonctionnelles et évolutives entre ces génomes. Pour une espèce, les relations entre les gènes peuvent être déduites d'après celles identifiées chez un organisme de référence et d'après les gènes prédits chez cette espèce, ce qui représente un gain de temps dans l'annotation de l'espèce. Les relations entre gènes peuvent s'étudier au niveau des interactions entre leurs produits d'expression, les protéines. Les différences observées entre espèces proches d'un point de vue phylogénique mettent ainsi en évidence les spécificités fonctionnelles de chacune de ces espèces. En revanche, les complexes protéiques évoluent peu, car ils sont soumis à une forte pression de sélection du fait de la complémentarité moléculaire nécessaire entre partenaires en interaction.

À ce jour, l'annotation d'un génome est un processus lent et complexe car elle procède de l'intégration de données et de résultats hétérogènes et en constante amélioration, de nature expérimentale et prédictive. Cette annotation permet de cibler les expérimentations en laboratoire qui, à leur tour, contribueront à un enrichissement de l'annotation.

L'annotation d'un génome est manuelle quand le choix du commentaire appartient à un expert biologique ; automatique lorsque le commentaire est imposé par l'analyse informatique. Le compromis est l'annotation semi-automatique où l'expert juge un commentaire proposé par un système informatique d'aide à la décision.

Cette thèse répond aux besoins spécifiques des trois niveaux de l'annotation manuelle de génomes apparentés. Un système d'annotation orientée génomique comparée a été mis en place, facilitant ainsi le travail des annotateurs et homogénéisant l'annotation. Puis l'application de règles de cohérence sur l'annotation permet d'en vérifier la qualité, détectant ainsi les erreurs d'annotation ou les spécificités des génomes annotés. Enfin l'apport d'une validation expérimentale de l'annotation pour la levure *Yarrowia lipolytica* permet, d'une part, d'obtenir des résultats originaux sur les relations protéiques de cette espèce et, d'autre part, de valider la démarche des règles de cohérence utilisées pour l'annotation génomique.

Cette thèse s'inscrit dans le projet d'annotation et de comparaison de génomes Génolevures [Souciet et al., 2000, Dujon et al., 2004] auquel collabore l'équipe de bio-informatique INRIA Futurs MAGNOME du LaBRI, en tant que responsable des analyses et ressources bio-informatiques. Impliquant une quinzaine de laboratoires et centres de recherche, le projet Génolevures est une étude de génomique comparative à grande échelle entre plusieurs

[1]Les anglo-saxons désignent par le terme "bioinformatics" la discipline spécifiquement consacrée à l'étude des séquences et des structures, et par le terme "biocomputing" le traitement sur ordinateur des données biologiques.

levures de la classe Hémiascomycète. Son objectif porte sur l'évolution moléculaire : évolution des gènes, familles fonctionnelles des protéines codées par ces gènes, mécanismes de réarrangement des chromosomes. Le travail préliminaire pour atteindre cet objectif repose sur l'annotation manuelle de quatre nouvelles espèces de levure. Cette annotation doit être de qualité car elle est à la base des études de comparaison génomique réalisées par la suite.

Problématiques et objectifs poursuivis

Disposer du résultat de l'annotation génomique est la condition *sine qua non* pour donner un sens biologique à l'ADN. Or l'obtention de ces résultats est soumise à des contraintes de temps et de qualité afin de les exploiter au plus tôt et de cibler les expérimentations biologiques. Ces dernières ayant un coût financier et temporel non négligeable pour les laboratoires de recherche, l'annotation doit donc être réalisée efficacement et être de grande qualité.

Contrainte de délais

La première contrainte dans un projet d'annotation est celle du *temps*. Les laboratoires de recherche, qu'ils soient publics ou privés, sont en situation de concurrence : la mise en valeur de leur travail se mesure par l'impact de leurs découvertes scientifiques, lors de leur publication et de leur communication à la communauté scientifique. Ces découvertes, et éventuellement les brevets, représentent ainsi une reconnaissance scientifique et/ou un intérêt financier, à condition d'être le premier (ou parmi les premiers) à les communiquer. La création de centres de séquençage à haut débit performants, tels que le Génoscope (France) ou le Broad Institute (États-Unis), rend abordable le séquençage de l'ADN génomique par les laboratoires de recherche biologique. Cette facilité leur permet de se lancer dans des projets de génomique comparée, domaine de recherche concernant l'évolution des espèces et la mise en relation du génome avec la physiologie d'un organisme. En diminuant le temps consacré à l'annotation, les laboratoires entrent plus vite au cœur de leurs recherches, espérant bénéficier de la primeur de leurs découvertes.

Jusqu'aux années 1990, les experts réalisaient l'annotation génomique de façon manuelle afin d'extraire l'information tant attendue à partir des séquences. Ils se partageaient le travail et se concertaient afin d'avoir une annotation la plus homogène possible. Encore actuellement, l'annotateur utilise des sources de données hétérogènes, aux formats variés, qu'il doit valider, croiser et intégrer, et ceci, pour chaque gène prédit. Le travail d'annotation est donc long et fastidieux. Grâce au développement des méthodes bio-informatiques d'aide à l'annotation, certaines tâches sont désormais automatisables. De ce fait, les laboratoires s'associent afin de disposer de plus de moyens humains, techniques et financiers pour effectuer les analyses. Les problèmes liés à l'organisation, la gestion et l'homogénéisation de ce travail communautaire sont résolus grâce à la mise en place d'outils de travail communs (portail internet d'annotation) et d'aide à la décision. L'annotation, manuelle ou semi-automatique, s'effectue alors plus rapidement, tout en donnant des résultats satisfaisants. À l'extrême, l'annotation automatique permet un gain de temps encore plus appréciable mais sa qualité est moindre (cf. § "Contrainte de qualité" ci-après).

En plus de cette contrainte de rapidité, les laboratoires de recherche publics disposent d'un court laps de temps pour tirer bénéfice de l'annotation génomique, car la politique des centres de séquençage publics est de mettre à disposition la séquence dans un temps imparti. Le séquençage d'un génome consiste en la lecture puis l'assemblage de petits fragments d'ADN génomique donnant des séquences, ou contigs, de tailles suffisantes pour commencer l'annotation. Lorsque le séquençage est réalisé par un centre de séquençage public, ce centre peut imposer la mise à disposition des séquences pour l'ensemble de la communauté scientifique passé un certain délai. Le laboratoire à l'initiative du séquençage voudra alors commencer l'annotation au plus tôt sans attendre l'assemblage final de la séquence génomique. Il faut toutefois noter que, plus l'assemblage est avancé, moins de zones significatives potentielles sont disjointes. Il est alors souhaitable que le système d'aide à l'annotation utilisé permette la transposition du travail d'annotation effectué pour une version d'assemblage du génome sur une autre, sans perte d'information. Le temps d'anticipation de l'annotation est autant de temps gagné pour les analyses réalisées ensuite.

Les systèmes d'aide à l'annotation doivent intégrer les méthodes couramment utilisées pour l'analyse de séquences. Ces méthodes se basent sur des méthodes *ab initio* (détection de caractéristiques biologiques et algorithmes probabilistes) et sur des méthodes d'homologie de séquence (recherche de similitudes significatives avec d'autres séquences biologiques connues). Ces méthodes portent, par exemple, sur la prédiction de gènes (GeneFinder [Green and Hillier, ults], GeneMark [Borodovsky and McIninch, 1993], *etc*), la comparaison de séquences (BLAST [Altschul et al., 1990], CLUSTAL W [Higgins et al., 1994]), la recherche de motifs (PROSITE [Sigrist et al., 2002], MEME [Bailey and Elkan, 1994]), la reconnaissance de repliements [Marin et al., 2002].

Si l'expert dispose des outils adéquats et des analyses pertinentes pour l'annotation, son effort d'annotation est concentré sur les cas complexes, diminuant ainsi le temps d'annotation.

Contrainte de qualité

La seconde contrainte à laquelle doit répondre un projet d'annotation est celle de la *qualité* de l'annotation : l'annotation génomique doit avoir une signification biologique cohérente. Cette qualité se définit, au niveau de l'annotation syntaxique, par sa *complétude*, *i.e.* l'identification exhaustive des gènes d'un organisme et, au niveau de l'annotation fonctionnelle et relationnelle, par l'*absence d'erreur* dans le commentaire relatif à chaque gène.

Les premiers organismes annotés avec une très grande qualité sont, en fait, tous des organismes modèles. Les organismes ont été choisis comme modèles par la communauté scientifique pour leur facilité de manipulation en tant qu'outil expérimental mais aussi pour l'étude d'un processus biologique particulier intéressant. Les expériences biologiques menées depuis le XIX^{e} siècle (voire le XVII^{e} siècle pour la levure) ont fourni des données en quantité importante et d'excellente qualité. Ces données précises et complètes (en ce qui concerne l'identification des gènes) ont permis une annotation de bonne qualité de ces organismes. Les organismes modèles les mieux connus sont : la bactérie colibacille[2], la levure de boulangerie[2] (ou levure de bière),

[2]La bactérie *Escherichia coli*, la levure *Saccharomyces cerevisiae*, le ver *Caenorhabditis elegans*, la souris *Mus musculus*, la mouche *Drosophila melanogaster*, le poisson zèbre *Danio rerio* et la plante *Arabidopsis thaliana*.

le ver nématode[2], la souris[2], la mouche du vinaigre[2], le poisson zèbre[2] et la "mauvaise herbe" arabette des Dames[2]. L'annotation d'organismes modèles sert à vérifier, par comparaison, la qualité de celle de nouveaux organismes qui leur sont proches d'un point de vue phylogénique.

Le premier critère qualitatif d'une annotation génomique est sa complétude, *i.e.* si l'annotation réalisée a identifié tous les éléments fonctionnels de l'organisme et si l'organisme ainsi décrit a un sens biologique cohérent. L'annotation génomique est effectuée en examinant les gènes candidats individuellement, sans avoir une vision globale du génome en tant que système fonctionnel cohérent. L'annotation résultante peut avoir manqué la détection d'éléments clefs, tels que ceux responsables de fonctions vitales pour l'organisme. Cette lacune a de graves répercussions pour sa compréhension, car cela biaise les expériences ultérieures et leur interprétation, engendrant une perte en temps et en argent.

Le second critère d'une annotation est son absence d'erreur. Lors du processus d'annotation, les premières sources d'erreurs sont les banques de données elles-mêmes, polluées par des erreurs d'origine humaine ou informatique [Galperin and Koonin, 1998, Jones et al., 2007]. L'annotation d'un génome nouvellement séquencé, basée sur les données provenant de ces banques, doit éviter autant que possible la propagation de ces erreurs.

Les secondes sources d'erreurs proviennent de séquences dont l'interprétation est ambiguë. C'est pourquoi l'obtention d'une annotation de qualité acceptable par annotation automatique est difficile, car cela suppose l'automatisation d'une tâche requérant le savoir-faire d'un annotateur humain biologiste, fort de sa connaissance du domaine et de son expérience. Lorsque plusieurs annotations distinctes sont possibles pour une même séquence, l'intervention humaine est réfléchie. L'expert est capable d'opter sciemment pour la seule annotation la plus en adéquation avec le fonctionnement de l'organisme, l'environnement du gène : il adapte la précision du commentaire afin que celui-ci reste juste. Par exemple, la transitivité d'annotation de séquence homologue en séquence homologue peut aboutir à un non-sens. Le gène annoté aura une annotation fonctionnelle prédite liée à une fonction métabolique absente de l'organisme étudié ou impliquant la présence d'un motif structural pourtant absent du gène.

La qualité d'une annotation peut être améliorée de plusieurs façons. Des règles de cohérence peuvent s'assurer que l'annotation respecte les contraintes biologiques (structure des éléments, fonctions vitales présentes ...). L'annotation d'un génome nouvellement séquencé doit également prendre en compte les informations connues pour cet organisme. De plus, la complétude de l'annotation d'un génome et sa justesse peuvent tirer profit de l'approche de génomique comparée. Ainsi la qualité de l'annotation peut se vérifier par comparaison avec l'annotation du plus proche des organismes modèles précédemment cités, afin que les différences observées révèlent uniquement les spécificités des organismes considérés. L'annotation des gènes orthologues[3] doit également être homogène.

À ce jour, les travaux de contrôle de cohérence d'annotation portent sur certains aspects de l'annotation, tels que la syntaxe de l'annotation (le commentaire attribué au gène), l'annotation fonctionnelle d'éléments biologiques similaires, les différences d'ontologie (vocabulaire

[3]Selon la définition de W. Fitch [Fitch, 2000], l'orthologie est la relation entre deux caractères homologues dont l'ancêtre commun est le plus récent ancêtre commun du taxon dont sont issues les séquences.

contrôlé) [Ashburner et al., 2000] observées entre les génomes. Mais ces travaux ne sont pas entièrement satisfaisants, car ils ne considèrent pas le génome comme une entité fonctionnelle.

Deux projets successifs, GASP [Reese et al., 2000] en 2000 et EGASP [Guigo et al., 2006] en 2006, ont évalué la qualité de l'annotation de logiciels d'annotation automatique, respectivement chez la drosophile et l'homme. Pour cela, les résultats d'annotation de ces logiciels étaient comparés et calibrés avec un jeu étalon de données annotées manuellement. Si certains des logiciels testés obtenaient de bons résultats, notamment pour le logiciel Exogean [Djebali et al., 2006], l'annotation automatique ne donne toujours pas d'aussi bons résultats que l'annotation manuelle ou semi-automatique. Toutefois, la meilleure façon de valider l'annotation reste l'expérimentation biologique. Quelle que soit la technique utilisée pour la validation de la structure (PCR, puces à ADN/ARN,...) et de la fonction d'un gène (mutation ou délétion du gène, marquage de la protéine,...), les laboratoires ne mobilisent pas ces moyens techniques et les moyens financiers associés dans le simple but de vérifier l'annotation. Ils visent un thème de recherche précis, tel que l'étude d'une famille de protéines ou un processus cellulaire donné.

Objectifs de la thèse et contribution

L'objectif de cette thèse est de proposer une réponse aux contraintes présentées précédemment, dans le cadre du projet Génolevures. Ainsi le travail présenté se restreint au contexte des organismes unicellulaires eucaryotes, et plus précisément la branche phylogénique des Hémiascomycètes dont fait partie l'organisme modèle *Saccharomyces cerevisiae* [Oliver et al., 1992, Goffeau et al., 1996].

Système d'aide à l'annotation

L'une des phases d'annotation de ce projet consistait en l'annotation manuelle de quatre génomes de levures nouvellement séquencés par les laboratoires impliqués. Afin de gagner du temps sur les analyses bio-informatiques et les expérimentations biologiques, les membres de Génolevures avaient deux souhaits. Le premier était de réaliser l'annotation manuelle (afin d'avoir une annotation de qualité) à l'aide d'un système informatisé accessible aux experts par internet afin d'avoir une annotation homogène. Ce système devait aussi intégrer les analyses bio-informatiques demandées par les biologistes.

Le second souhait était de commencer l'annotation avant la fin de l'assemblage des séquences réalisé par le Génoscope afin de disposer de plus de temps pour mener à bien les analyses. Comme nous l'avons vu précédemment, le temps est un facteur important dans ce domaine de la recherche.

La première partie de ce travail de thèse porte sur l'adaptation et l'amélioration du logiciel d'aide à l'annotation CAAT-Box dans le cadre de la génomique comparée. Le logiciel CAAT-Box ("Contig-Assembly and Annotation Tool-Box") [Frangeul et al., 2004], développé en 2000 à l'Institut Pasteur, permet le transfert de l'annotation des gènes prédits d'une version de génome sur une autre, grâce à un système de gestion de version de séquences et de "remapping". Il permet également l'intégration aisée d'informations "à la carte" sur le gène grâce à un

système simple de nommage de fichiers. Mais ce système a été développé pour l'annotation des organismes procaryotes, dont le génome a une structure plus simple que celle des eucaryotes. Nous avons par conséquent modifié le système CAAT-Box afin de prendre les comparaisons entre génomes et les caractéristiques biologiques présentes chez les eucaryotes, telle que la détection des introns, éléments géniques ne codant pas de séquence protéique.

Cohérence de l'annotation

Le projet Génolevures nous offre un contexte favorable pour contrôler la qualité d'une annotation génomique. En effet, les espèces étudiées sont les levures hémiascomycètes dont fait partie l'organisme modèle *S. cerevisiae*. Les différences d'annotation observées entre celle de *S. cerevisiae* et celles des autres levures, dont la complétude est à vérifier, devraient mettre en évidence les fonctions où la spéciation a opérée. Nous avons ainsi développé des règles de cohérence afin de vérifier la qualité de ces annotations génomiques.

Ces règles vérifient le respect des contraintes biologiques admises par la communauté scientifique et se basent sur la logique. Nous avons défini trois niveaux d'application de ces règles. Le premier ensemble de règles s'applique au niveau de l'élément génique. Par exemple, elles vérifient la présence des deux codons, initiateur et terminateur, sur la séquence codante prédite et le respect du cadre de lecture après épissage des introns. Le second ensemble de règles s'applique au niveau du voisinage d'un élément, jusqu'à considérer le chromosome entier. Ces règles vérifient, par exemple, que les éléments fonctionnels ne se chevauchent pas. Puis le troisième ensemble de règles s'appliquent au niveau du génome. Par exemple, ces règles vérifient la présence des gènes décrits dans la littérature, la présence des fonctions cellulaires vitales. Nous avons défini ces règles à partir de la littérature scientifique, d'analyses de séquences et d'études bio-informatiques. Ces règles de cohérence présentent trois avantages. D'abord, elles définissent si le génome est complet et a un sens biologique correct. Elles définissent ensuite le niveau informatif d'une différence observée avec le génome de référence, *i.e.* si celle-ci est due à une erreur d'annotation ou à une caractéristique de l'espèce considérée. Enfin, ces règles de cohérence, associées à des règles d'annotation, permettent une meilleure gestion du travail des annotateurs et garantissent ainsi une annotation de grande qualité, qu'elle soit manuelle ou automatique. Ces règles sont en partie implémentées et intégrées à la plate-forme d'annotation mise en place pour la seconde phase d'annotation de génomes dans le projet Génolevures. Elles sont adaptables en fonction de l'état de connaissance au moment où elles s'appliquent. De plus, la méthode est suffisamment généraliste pour s'appliquer à des branches phylogéniques autres que celle des Hémiascomycètes.

Validation expérimentale : étude des complexomes

Nous nous sommes également intéressés, en parallèle, à l'étude expérimentale du complexome (l'ensemble des complexes protéiques d'une cellule) chez la levure *Y. lipolytica*. L'étude des interactions protéine-protéine (IPP) existantes au sein d'un organisme ou d'une cellule permet d'approfondir les connaissances sur ses mécanismes biologiques et d'établir les relations existantes entre ses protéines. Cette étude nous a permis, par la même occa-

sion, de valider de façon expérimentale la qualité de l'annotation réalisée par le consortium Génolevures.

Pour cette validation expérimentale de la qualité de l'annotation, nous avons utilisé la méthode expérimentale de l'électrophorèse bi-dimensionelle en bleu natif et SDS [Camacho-Carvajal et al., 2004]. Cette technique permet d'observer la composition des complexes protéiques détectés à l'échelle de la cellule en condition non dénaturante.

Le contrôle expérimental de cette technique est l'étude du complexome chez *S. cerevisiae*. En effet, les complexes protéiques de *S. cerevisiae* ont fait l'objet de nombreuses études biologiques à grande échelle et sont bien caractérisés [Uetz et al., 2000, Ito et al., 2001, Gavin et al., 2002, Ho et al., 2002].

Dans un premier temps, la détection expérimentale de protéines prédites pour *Y. lipolytica* confirme l'annotation syntaxique réalisée. Dans un second temps, la mise en évidence de complexes protéiques chez *Y. lipolytica* valide l'induction des règles de cohérence et leur signification. Par comparaison avec les complexes de *S. cerevisiae*, nous pouvons prédire les interactions protéine-protéine dans les différences espèces de levures étudiées lors du projet Génolevures. L'étude des différences observées peut ainsi révéler la spécificité de chacune de ces espèces, à condition que l'annotation soit de bonne qualité.

Nous avons ainsi identifié en partie les complexes protéiques pour *Y. lipolytica* de façon expérimentale. Cette vérification est partielle en raison de la sensibilité de détection de la technique employée.

Le choix s'est porté sur *Y. lipolytica* pour trois raisons. Tout d'abord, nous disposerons ainsi de nouvelles données biologiques sur cette espèce. Ensuite, *Y. lipolytica* est la levure la plus distante de *S. cerevisiae* parmi les levures étudiées dans le projet Génolevures. Les informations expérimentales obtenues nous permettent ainsi d'encadrer la comparaison des levures hémiascomycètes. Enfin, nous avons choisi *Y. lipolytica* car cette levure change de métabolisme selon son environnement de croissance. Nous pouvons ainsi observer des différences d'expression génique pour cette espèce.

Organisation de la thèse

Après ce premier chapitre introductif, nous présentons dans le deuxième chapitre les notions de biologie utiles à la compréhension du domaine dans lequel ce travail de thèse s'est réalisé. Il traite notamment des interactions protéine-protéine, de leurs intérêts et de leurs techniques d'étude.

Puis nous définissons, dans le troisième chapitre, l'annotation génomique. Nous abordons l'annotation relationnelle en discutant des différentes techniques expérimentales pour l'étude des complexes protéiques dans une cellule. Nous analysons, ensuite, les approches mises en œuvre à ce jour dans le cadre de l'annotation génomique, en considérant les aspects de rapidité et de qualité, ainsi que les méthodes sur son contrôle et sa validation expérimentale. Nous indiquons à quel degré elles satisfont ces contraintes.

Le quatrième chapitre présente notre contribution en terme de règles de cohérence appliquées pour le contrôle de la qualité de l'annotation. Nous détaillons ces règles de logique à

leurs différents niveaux d'action : de l'élément codant pris individuellement à l'ensemble du génome. Nous présentons aussi le ou les intérêt(s) particulier(s) de chacune de ces règles.

Dans le cinquième chapitre, nous exposons la mise en œuvre de ces règles de cohérence dans le cadre du projet Génolevures. Nous présentons aussi la stratégie d'annotation manuelle de ce projet et la plate-forme d'annotation mise en place. Puis nous présentons l'implémentation et les résultats de l'utilisation de certaines de ces règles sur les divers organismes annotés par Génolevures.

Dans le dernier chapitre, nous présentons l'étude du complexome chez *Y. lipolytica*, en comparant les prédictions avec leur validation expérimentale. Nous définissons notre technique de séparation et d'identification de protéines.

Enfin, nous concluons sur le travail réalisé pour l'annotation d'un génome et le contrôle de sa cohérence par des méthodes bio-informatiques et expérimentales. Nous terminons tout naturellement par les perspectives de ce travail qui portent, entre autres, sur l'intégration des règles de cohérence dans la plate-forme d'annotation utilisée actuellement pour une autre phase du projet Génolevures.

Chapitre 2

Notions élémentaires de biologie

Sommaire

Cette introduction à la biologie ne se veut pas exhaustive mais suffisamment détaillée pour comprendre les données et les problèmes discutés dans ce document. Nous présentons les notions élémentaires de biologie pour une cellule telle que la levure. En biologie, la levure étant l'organisme de référence pour le domaine des Eucaryotes, ces notions ont donc un caractère généraliste.

En premier lieu, nous définissons la cellule eucaryote et ses composants. Puis nous présentons les trois grands mécanismes cellulaires qui régissent l'activité de la cellule : la réplication, la transcription et la traduction. Nous présentons ensuite la diversité des interactions entre molécules présentes dans une cellule, et en particulier celles entre protéines. Ces interactions sont essentielles au bon fonctionnement de la cellule. L'interaction reposant sur une complémentarité entre ses partenaires mis en jeu, ceux-ci sont soumis à une forte pression de sélection. La comparaison des interactions protéiques témoigne ainsi de l'évolution des espèces.

2.1 Organismes vivants

2.1.1 Classification en règnes

A l'heure actuelle, la classification phylogénique, basée sur l'évolution des espèces et la notion d'ascendance commune, divise le monde vivant en deux empires : les Procaryotes et les Eucaryotes.

Les Eucaryotes (du grec *eu*, vrai et *karuon*, noyau) se différencient des Procaryotes (du grec *pro*, avant et *karuon*, noyau) par la présence d'un noyau qui renferme leurs molécules d'ADN linéaires (une molécule circulaire principale chez les Procaryotes), d'un cytosquelette et des mitochondries.

D'après la classification de Woese [Woese et al., 1978, Woese et al., 1990], classification la plus couramment utilisée, le monde du vivant se compose de trois domaines : Eucaryotes, Archéobactéries et Bactéries (ou Eubactéries). Les domaines des Archées et des Bactéries forment l'empire des Procaryotes, unicellulaires, opposé à celui des Eucaryotes. Le domaine des Eucaryotes regroupe quatre règnes :

- les Protistes, unicellulaires,
- les Champignons, pluricellulaires ou unicellulaires,
- les Végétaux, pluricellulaires,
- les Animaux, pluricellulaires.

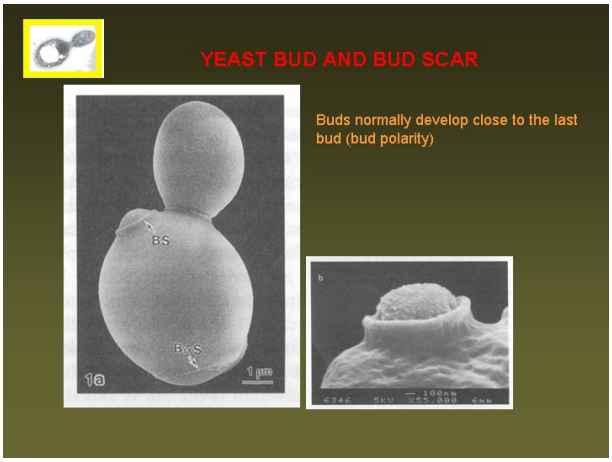
2.1.2 Caractéristiques communes

Les organismes vivants eucaryotes et procaryotes partagent plusieurs caractéristiques. Leur(s) cellule(s) est (sont) entourée(s) d'une membrane bicouche lipidique, dite membrane plasmique, contenant une substance riche en eau, le cytoplasme. Ces organismes se multiplient, assurant ainsi leur pérennité. Le support de leur information génétique est la molécule d'ADN, qui est répliquée à l'identique au fil des générations. Ces organismes vivants possèdent les moyens d'utiliser l'information contenue dans l'ADN pour fabriquer (synthétiser) les éléments constitutifs de leur(s) cellule(s) : protéines, lipides, glucides et métabolites.

Pour la compréhension du contexte de cette thèse, nous nous contentons de développer les notions de biologie pour la cellule eucaryote.

FIG. 2.1 – Levure *S. cerevisiae* bourgeonnante.

1a : levure bourgeonnante, avec les cicatrices laissées par les précédentes cellules filles (BS : "Birth Scars"). 1b : grossissement d'une cicatrice de naissance (source : [Feldmann, 2005]).



2.1.3 La cellule

La cellule est l'unité structurale, fonctionnelle et reproductrice constituant tout ou partie de l'être vivant. Une cellule, selon les éléments qui la composent, définit son appartenance à l'un des règnes vivants.

2.1.3.1 Caractéristiques du modèle de cellule eucaryote *S. cerevisiae*

La cellule eucaryote modèle en biologie est *Saccharomyces cerevisiae* plus communément appelée levure de boulangerie ou levure de bière. Sa facilité d'observation, de culture, de manipulation en a fait un véritable outil de prédilection pour le biologiste, et ce, dès la fin du XVII^{e} siècle. Ainsi, de nombreux processus cellulaires et moléculaires communs à toutes les cellules ont été découverts grâce à *S. cerevisiae*.

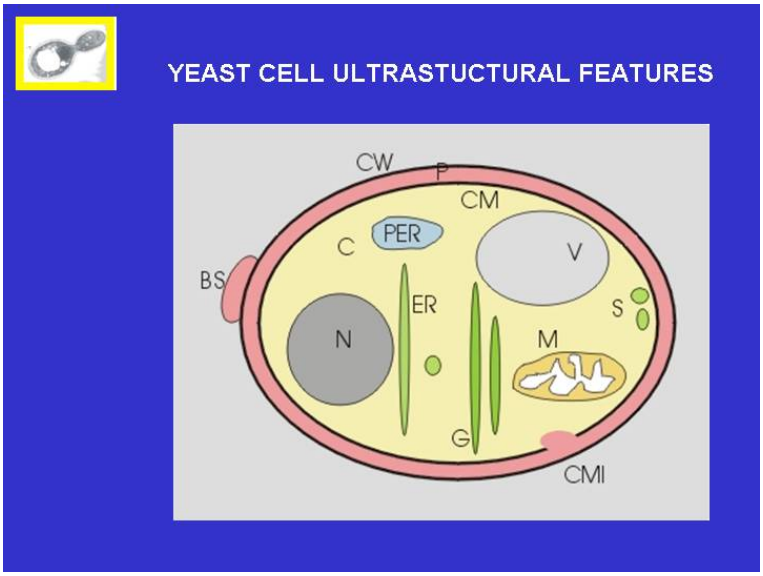
La levure *S. cerevisiae* (cf. fig. 2.1) pousse sur milieu nutritif solide ou liquide contenant une source de carbone, telle que le glucose, à une température entre 15°C et 37°C. Par défaut, en présence de sucre et d'oxygène, la levure puise son énergie, de préférence, à partir du métabolisme de la fermentation alcoolique : le glucose est alors dégradé et converti en gaz carbonique et éthanol (alcool éthylique). En l'absence de sucre, elle passe alors au métabolisme de la respiration.

La cellule de levure a une taille de 5 à 10 μm (dans son diamètre le plus large) et elle est de forme ovoïde, assurée par sa paroi cellulaire et son squelette interne (le cytosquelette).

Au cours de sa vie, la cellule mère donne naissance par bourgeonnement (cf. fig. 2.1) à environ 20 cellules filles. Le temps d'une génération est de 2 à 3h à 28°C.

FIG. 2.2 – Organisation d'une cellule de levure.

CW = paroi cellulaire, CM = membrane plasmique, P = périplasme, C = cytoplasme (cytosol), N = noyau, ER = réticulum endoplasmique, G = appareil de Golgi, V = vacuole, S = vésicules de sécrétion, M = mitochondrie, BS = bourgeon, CMI = invagination de la membrane plasmique, PER = peroxysome. Les protéasomes ne sont pas représentés (source : [Feldmann, 2005]).



2.1.3.2 Éléments structuraux de la cellule

La cellule eucaryote contient de grandes surfaces membranaires. Toutes les fonctions cellulaires sont compartimentées et réalisées par des organites spécialisés, structures délimitées par une membrane.

La figure 2.2 représente l'organisation d'une cellule de levure avec ses différents organelles et compartiments. Nous allons décrire, à l'aide de ce schéma, ces divers éléments (l'abréviation entre parenthèses reprend celle utilisée sur le schéma).

L'enveloppe cellulaire

L'**enveloppe cellulaire** contrôle l'osmolarité (concentration en éléments, ioniques ou non) et la perméabilité de la cellule. Elle se compose de trois éléments (de l'extérieur vers l'intérieur) aux fonctions précises :

- la **paroi cellulaire** (CW) : formée principalement de polysaccharides, elle donne sa forme à la cellule,
- le **périplasme** (P) : compartiment contenant principalement des protéines sécrétées qui ne traversent pas la paroi et des enzymes qui hydrolysent les substrats nutritifs qui ne traversent pas la membrane plasmique
- la **membrane plasmique** (CM) :formée d'une double couche phospholipidique, avec

du cholestérol, elle est traversée par des protéines (récepteurs aux molécules extérieures, canaux, transporteurs) ; elle contrôle la perméabilité de la cellule.

Le noyau

Délimité par l'enveloppe nucléaire (une double membrane bilipidique) traversée par les pores nucléaires, le **noyau** (N) renferme le matériel génétique présent sous la forme de molécules d'ADN linéaires, les chromosomes. Selon les espèces, plusieurs éléments non chromosomiques peuvent également être trouvés dans le noyau tels que le(s) plasmide(s), une petite molécule d'ADN circulaire. Par exemple, *S. cerevisiae* possède 16 chromosomes et un plasmide appelé 2μ du fait de sa taille.

Le système sécrétoire et la vacuole

Le **réticulum endoplasmique** (ER) est constitué d'un réseau de membranes étendu, en continuité avec l'enveloppe nucléaire et en relation avec d'autres compartiments tels que l'appareil de Golgi. Les protéines synthétisées à l'extérieur du réticulum pénètrent dans celui-ci pour y subir des modifications (ajout de sucres) et des repliements afin d'être fonctionnelles. Le réticulum endoplasmique est également impliqué dans l'assemblage et le transport des protéines destinées aux membranes et à la sécrétion.

L'**appareil de Golgi** (G) est constitué d'un empilement de plusieurs petits sacs membranaires en forme de disque, en relation avec le réticulum endoplasmique grâce à un système de vésicules de transport. Il régule le nombre de vésicules allant à la membrane plasmique et participe ainsi au renouvellement membranaire. Il est le lieu des modifications post-traductionnelles des protéines (maturation des protéines, ajout de chaînes sucrées, de phosphate et sulfate), nécessaires à l'adressage correct de ces protéines dans la cellule ou vers la sécrétion.

La **vacuole** (V) est l'organelle clef de la levure. Cette vésicule, entourée d'une membrane phospholipidique simple, au contenu acide, remplie d'enzymes de dégradation pour de grandes molécules (comme les protéines) devenues inutiles, nuisibles ou dégradées, et permettant ainsi à la cellule de récupérer les molécules de base.

De même que des vésicules assurent le transport des protéines entre le RE, l'appareil de Golgi, la vacuole et leur exocytose à l'extérieur de la cellule, des vésicules d'endocytose (ou endosomes) se forment par déformation de la membrane plasmique lors de l'absorption de molécules complexes, et se combinent avec la vacuole pour leur dégradation.

La mitochondrie

Les **mitochondries** (M) sont les "centrales énergétiques" (ou les "poumons") de la cellule car les deux dernières étapes de la respiration (en présence d'oxygène), que sont le cycle de Krebs et la chaîne de transport d'électrons, s'y déroulent. La première étape, la glycolyse (dégradation du glucose), se déroule dans le cytoplasme. L'énergie des molécules organiques issues de la glycolyse est convertie en énergie directement utilisable par la cellule sous la forme de molécules d'ATP. En l'absence d'oxygène, la cellule utilise la fermentation dans le cytoplasme pour produire l'énergie nécessaire à son fonctionnement mais ce système a un rendement beaucoup moins efficace.

La mitochondrie possède son propre génome, codant une petite partie (28 gènes chez la levure) des protéines mitochondriales, le reste étant codé par le génome nucléaire.

Lieux de dégradation des molécules

Localisés dans le cytosol, le noyau et le réticulum endoplasmique, les **protéasomes** sont des complexes protéiques, en forme de tonneau, responsables de la dégradation des protéines, recyclant ainsi les constituants cellulaires. Ils dégradent les protéines mal repliées, dénaturées ou inutiles en les découpant en peptides de 7 à 9 acides aminés qui seront complètement hydrolysés en dehors du protéasome.

Les **peroxysomes** (PER), entourés d'une membrane simple, protègent la cellule car ils sont responsables de l'élimination des radicaux libres produits par l'oxygène dans la cellule. Ils sont aussi le lieu de la dégradation des acides gras.

L'ensemble des organites "baignent" dans le cytosol, ou cytoplasme (cf. fig. 2.2), gel visqueux intracellulaire riche en eau. Le déplacement d'une particule (protéine, petite molécule, vésicule de transport) se fait le long de microtubules (cf. fig. 2.2) (vus comme un réseau ferroviaire). Ces microtubules jouent également le rôle de squelette pour la cellule, responsable de sa forme et de son intégrité lors de variation de pression osmotique.

2.2 Molécules du vivant

Le gène est l'unité de base de l'information génétique d'un organisme. Il constitue une partie de la molécule d'ADN et est exprimé sous la forme d'une molécule d'ARN qui, à son tour, peut coder ou non une molécule d'acides aminés, la protéine. Ces deux dernières molécules sont les molécules exprimées participant au fonctionnement de la cellule.

2.2.1 ADN

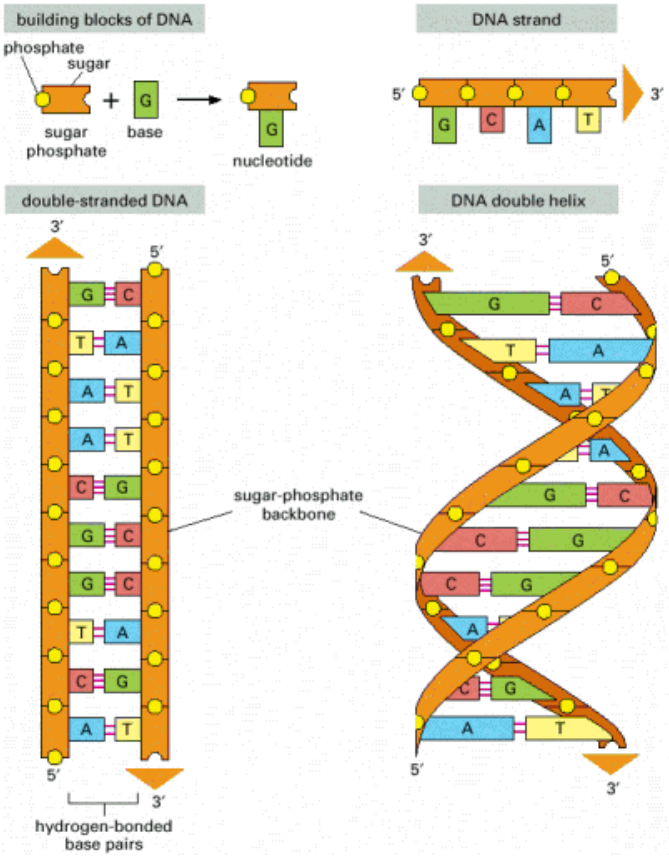
La molécule d'acide désoxyribonucléique (ADN) (cf. fig. 2.3) se compose de deux brins enroulés l'un autour de l'autre (molécule bicaténaire ou double brin), lui donnant une structure de double hélice (cf. fig. 2.3 "DNA double helix"), et orientés en sens opposé l'un par rapport à l'autre (molécule antiparallèle) (cf. fig. 2.3 "double strand DNA").

Sa structure a été mise en évidence en 1953 par James Watson et Francis Crick[1] [Watson and Crick, 1953b], plusieurs années après que sa composition chimique fût connue.

Chaque brin d'ADN est composé de l'enchaînement orienté (cf. fig. 2.3 "DNA strand") de nucléotides, eux-mêmes composés d'un sucre (le désoxyribose), d'un phosphate et d'une base azotée (cf. fig. 2.3 "building blocs of DNA"). Il existe quatre bases azotées au niveau de l'ADN : l'adénine (notée `A`), la thymine (notée `T`), la cytosine (notée `C`) et la guanine (notée `G`).

[1] Avec la précieuse aide de Maurice Wilkins et Rosalind Franklin, dont le rôle fut dévalorisé par les trois chercheurs dès sa mort prématurée en 1958. Les trois hommes reçurent le prix Nobel de médecine en 1962 pour leurs travaux sur l'ADN.

FIG. 2.3 – Structure de la molécule d'ADN.
(source : [Alberts et al., 2002])



Les molécules d'ADN d'une cellule (les chromosomes) ont une longueur de plusieurs millions de nucléotides.

Par convention, une séquence d'ADN (monocaténaire) est représentée en fonction de la base présente sur le nucléotide et dans le sens 5' vers 3', selon la numérotation des atomes de carbone de sucre à partir desquels se font les liaisons covalentes (*i.e.* de forte énergie). Une molécule double brin s'écrit comme suit :

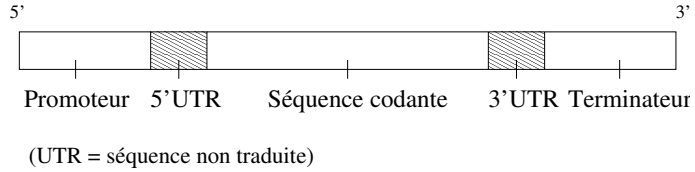
`5' GATACAATGAAATGTGTATC 3'` brin sens ou brin "+" ou brin Watson
`3' CTATGTTACTTTAGACATAG 5'` brin anti-sens ou brin "-" ou brin Crick

La longueur d'une séquence se mesure en nucléotides *nt* ou *b* (pour bases) si elle est simple brin, ou *pb* (pour paire de bases) si elle est double brin.

Les deux brins d'ADN s'apparient entre eux par des liaisons non covalentes (*i.e.* de faible énergie) grâce à la complémentarité des nucléotides deux à deux : `A` se lie à `T`, `C` se lie à `G`. Les paires de bases `AT` et `GC` sont les seules possibles de façon stable (cf. fig. 2.3 "double-strand DNA").

L'ADN contient les informations déterminant le développement et le fonctionnement de l'organisme. Il est le support de l'hérédité car il transmet cette information génétique à la

FIG. 2.4 – Structure d'un gène.



descendance en se répliquant. Cette information peut se modifier lentement au cours du temps, permettant l'évolution des espèces par pression sélective naturelle.

L'ensemble des molécules d'ADN d'une cellule constitue son génome. Ainsi le génome est spécifique d'un organisme et constant dans le temps. Seules certaines régions d'ADN ont une signification particulière : gènes, séquences répétées, promoteur.... Un abus de langage restreint le génome à l'ensemble des régions codantes, ses gènes. Ces régions sont dites codantes car l'information qu'elles contiennent est exprimée en éléments fonctionnels (protéines ou petits ARN) par le biais de molécules d'ARN. Le gène est l'unité fonctionnelle de la molécule d'ADN. Un gène (cf. fig. 2.4) codant une protéine se compose de plusieurs éléments :

- le promoteur (à gauche sur le schéma) : il régule l'expression du gène et contient les séquences reconnues par le complexe de l'ARN polymérase (cf. 2.3.2),
- les régions 5'et 3' non traduites (UTR, "UnTranslated Region") (de part et d'autre de la séquence codante sur le schéma),
- la séquence codante (CDS, "Coding DNA Sequence") : l'information génétique au sens strict,
- le terminateur (à droite sur le schéma) : il déclenche le décrochage de l'ARN polymérase lors de la transcription.

2.2.2 ARN

L'acide ribonucléique (ARN) est une molécule proche de la molécule d'ADN dans sa structure et sa fonction. L'ARN diffère cependant de l'ADN en quatre points : l'ARN est simple brin ; le sucre est le ribose (et non le désoxyribose de l'ADN) ; la thymine est remplacée par l'uracile (le nucléotide est noté `U`) ; les séquences d'ARN sont courtes (50 à 50 000 nt) par rapport à l'ADN.

Les ARN ont deux fonctions distinctes, chacune exécutée par un ou plusieurs types d'ARN spécifiques :

- l'ARN est un support intermédiaire de l'information génétique : ce rôle est assuré par l'ARN messager (ARNm) qui transmet ainsi l'information du gène à l'extérieur du noyau et qui servira à la synthèse des protéines,
- l'ARN a une fonction intrinsèque : il participe à certaines fonctions cellulaires telles que la traduction (les ARN de transfert ARNt, les ARN ribosomiques ARNr) ou l'épissage (les petits ARN nucléaires ARNsn), il sert de guide aux protéines en ciblant une séquence nucléique spécifique (les petits ARN nucléolaires ARNsno).

L'ensemble des molécules d'ARN d'un organisme est appelé le transcriptome. Celui-ci varie au cours du temps, selon l'environnement et, pour les organismes pluricellulaires, selon le type cellulaire ou le tissu cellulaire.

2.2.3 Protéine

Une protéine est constituée d'un enchaînement d'acides aminés, l'unité de base de cette séquence, aux propriétés physico-chimiques différentes des molécules nucléiques. Il existe 20 acides aminés à l'état naturel, représentés chacun par une lettre (`M` pour la méthionine, `R` pour l'arginine...). Chaque acide aminé a une structure particulière, chargée électriquement (positif/ négatif) ou non, responsable de la conformation d'une protéine. Les acides aminés sont regroupés selon des propriétés communes : acides aminés acides, basiques, neutres... Ces acides se lient entre eux par des liaisons covalentes, dites liaisons peptidiques. Quatre niveaux de structures caractérisent une protéine :

- la structure primaire (structure 1D) : la succesion linéaire des acides aminés,
- la structure secondaire (structure 2D) : le repliement local de la séquence protéique, dont les trois principales catégories ont des noms évocateurs de leur forme : hélices, feuillets et coudes,
- la structure tertiaire (structure 3D) : le repliement de la séquence dans l'espace, permettant à la protéine d'être fonctionnelle (une protéine qui a perdu sa structure 3D est dite dénaturée),
- la structure quaternaire (structure 4D) : l'association de plusieurs séquences polypeptidiques entre elles par liaisons non covalentes, chacune des séquences est appelée monomère ou sous-unité, l'association de plusieurs chaînes est désignée par oligomère ou protéines multimériques, ou par extension complexe protéique.

L'ensemble des protéines d'un organisme est appelé le protéome. Comme pour le transcriptome, le protéome varie au cours de la vie de la cellule, selon son environnement et, pour les organismes pluricellulaires, selon le type cellulaire ou le tissu cellulaire.

Une protéine se caractérise par ses principales propriétés physico-chimiques telles que sa longueur exprimée en acides aminés *aa*, son poids moléculaire exprimé en Dalton *Da* [2], sa charge électrique, son point isoélectrique[3] mais aussi par sa composition en domaines fonctionnels.

Un domaine fonctionnel est une région peptidique de la protéine responsable, en tout ou partie, de la fonction et de la structure de la protéine. À chaque type de domaine correspond un motif peptidique. Chaque protéine peut être constituée de plusieurs domaines.

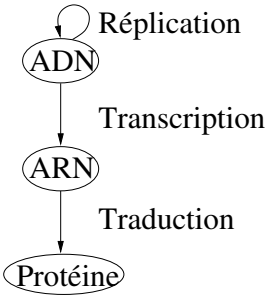
[2] Le poids moyen d'un acide aminé est d'environ 110 Da. Le Dalton est égal à $1,66.10^{-27}$ kg : il équivaut à $\frac{1}{12}^{e}$de la masse d'un atome de carbone 12, et exprime la masse d'un atome d'hydrogène. Cette unité de poids atomique porte le nom de son inventeur John Dalton (1766-1844), chimiste et physicien, mais aussi le premier à décrire dans une revue scientifique une anomalie de la vision dont il souffrait : le daltonisme.

[3] Le point isoélectrique ou pI, est le pH (potentiel Hydrogène : activité qui mesure l'acidité ou la basicité d'une solution) du milieu tel que la charge électrique globale de la protéine est nulle. Le pI intervient dans les techniques de séparation des protéines.

Le fonctionnement cellulaire repose sur trois mécanismes fondamentaux, irréversibles chez les eucaryotes en conditions naturelles et sans apport extérieur (tel que le fait un virus) : la réplication, la transcription et la traduction. Chacun de ces mécanismes intervient au niveau de ces molécules :

- la réplication permet le maintien de la molécule d'ADN lors des divisions cellulaires, *i.e.* la transmision de l'information génétique,
- la transcription permet l'expression de l'information contenue sur l'ADN en ARN,
- la traduction permet l'expression de l'information contenue sur les molécules d'ARN en protéines.

FIG. 2.5 – Les trois dogmes en biologie.



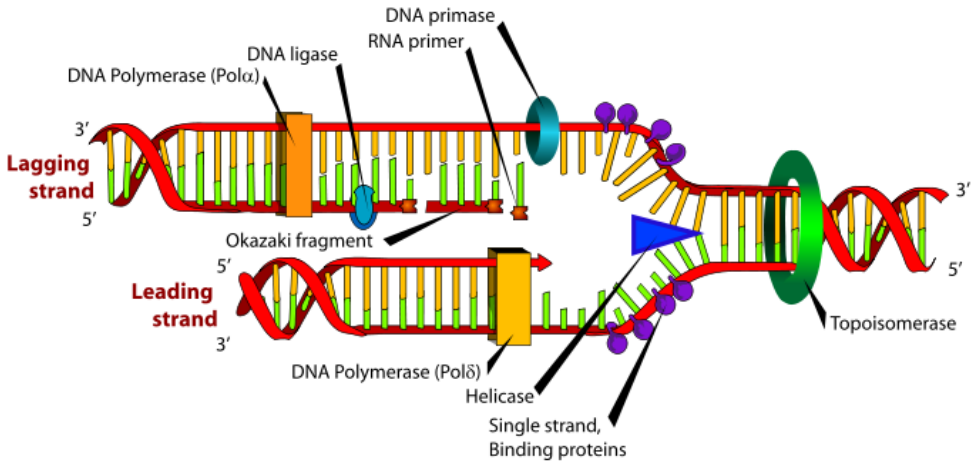
2.3 Mécanismes cellulaires

2.3.1 Réplication de l'ADN

La réplication (cf. fig. 2.6) ne concerne que les molécules d'ADN : elle est donc localisée dans le noyau et les mitochondries de la cellule et les chloroplastes dans le cas des cellules végétales. La réplication permet à l'ADN de se reproduire et de se transmettre à la cellule fille lors de la division cellulaire. Elle est assurée par plusieurs complexes enzymatiques, dont les ADN polymérases α et δ (chacune spécifique d'un brin d'ADN) (cf. fig. 2.6, Polα et Polδ), composés de nombreuses protéines provenant du cytoplasme et assemblées dans le noyau. Le double brin d'ADN est séparé par une protéine particulière, l'hélicase (cf. fig. 2.6, à droite), au niveau de la fourche de réplication. Chacun de ces brins, appelé brin parental, sert alors de matrice pour la synthèse d'un nouveau brin, ou brin néoformé, qui lui est complémentaire, formant ainsi une nouvelle molécule double brin.

La réplication est un processus bi-directionnel. En effet, les nouveaux brins sont synthétisés en même temps, de façon différente due à l'orientation du brin. La synthèse d'un brin nucléotidique, ADN ou ARN, se fait dans le sens 5' vers 3' pour des raisons biochimiques. Ainsi, le brin d'ADN 5'-3' néoformé est synthétisé de façon continue, tandis que le brin d'ADN néoformé 3'-5' est synthétisé par plusieurs fragments (les fragments d'Okasaki, cf. fig. 2.6) qui

FIG. 2.6 – Réplication de l'ADN.
(source : Mariana Ruiz Villarreal).



sont ligaturés par l'ADN ligase, au fur et à mesure de la progression du complexe de l'ADN polymérase α. La réplication est également qualifiée de processus semi-conservateur car chaque molécule d'ADN est composée d'un brin parental et d'un brin néoformé.

La réplication de l'ADN se doit d'être un processus très fiable puisqu'elle opère sur des millions de bases et qu'elle garantit l'intégrité de l'information génétique. Le complexe enzymatique de la réplication possède un système de détection et réparation en cas d'erreur. Par ailleurs, la cellule possède un système de réparation de l'ADN indépendant qui contrôle l'état de la molécule d'ADN (mauvais appariement nucléotidique, absence de nucléotide, cassure d'un brin...). Malgré ces contrôles, une mutation génétique peut persister, entraînant des effets allant d'un effet nul à la mort cellulaire si elle nuit à l'expression correcte d'une protéine vitale pour la cellule, en passant par une prolifération anarchique de la cellule (ce qui, chez les eucaryotes supérieurs, provoque des cancers).

Grâce à la haute fidélité de la réplication, l'information génétique persiste pour les générations cellulaires à venir mais elle n'a de valeur que dans son expression sous forme d'ARN et/ou de protéine. Cette expression se fait en deux étapes (cf. fig. 2.7) : la transcription, obligatoire pour l'expression de toute information génétique, et la traduction, nécessaire pour la synthèse des protéines seulement.

2.3.2 Transcription de l'ADN en ARN

La transcription (cf. fig 2.8), comme la réplication, est également localisée dans le noyau, l'ADN ne sortant jamais du noyau. La transcription consiste en la recopie d'une région d'ADN en une séquence d'ARN équivalente. Ce mécanisme de recopie est assurée par un complexe enzymatique multiprotéique composé d'une ARN polymérase et de facteurs de transcription. Il existe trois ARN polymérases différentes selon le type de molécules d'ARN synthétisées : les ARN polymérases de type 1 et 3 pour les ARN non codants, et l'ARN polymérase de type 2

FIG. 2.7 – Du gène à la protéine.
(source : Guillaume Baukiau)

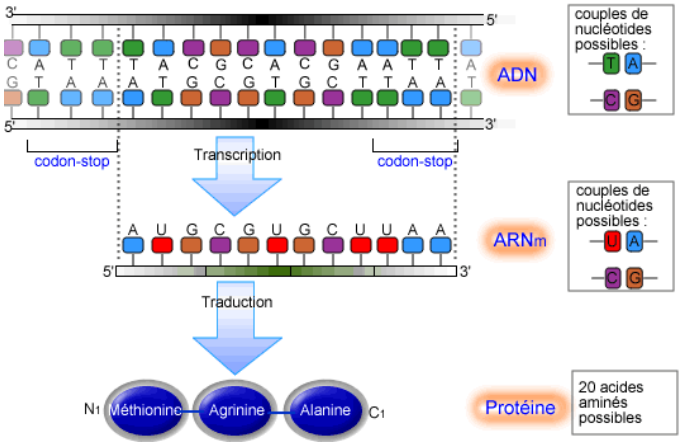
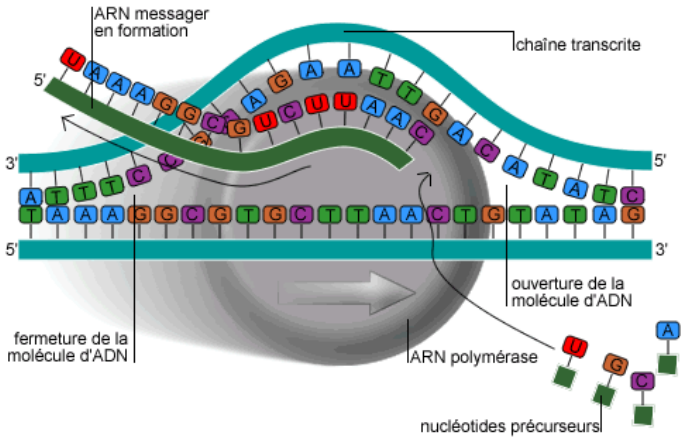


FIG. 2.8 – Transcription de l'ADN en ARN.
(source : Guillaume Baukiau)



pour les ARNm.

Selon sa fonction, la séquence d'ARN synthétisée pourra sortir du noyau par les pores nucléaires.

Quels que soient les types d'ARN synthétisés et d'ARN polymérase, le mécanisme est le même et se déroule en trois étapes :

- l'initiation : l'ARN polymérase et ses cofacteurs reconnaissent trois régions spécifiques du promoteur du gène sur lesquelles ils se fixent,
- l'élongation : le complexe ARN polymérase effectue la synthèse de la molécule d'ARN à l'identique (excepté le changement du `T` en `U`) de la séquence d'ADN codante, en utilisant le brin non codant (brin complémentaire orienté 3'-5') comme matrice et en se déplaçant progressivement le long de l'ADN (cf. fig 2.8) et

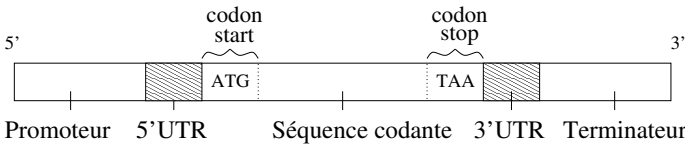
– la terminaison : le complexe de transcription détecte une région spécifique du gène, le terminateur, qui déclenche le décrochage du complexe enzymatique.

Transcription d'un gène codant une protéine

Chez les eucaryotes comme chez les procaryotes, un gène codant une protéine a une structure particulière qui lui permet d'être traduit en protéine (cf. fig. 2.9). Sa séquence codante est délimitée par deux triplets de nucléotides particuliers :

– à l'extrémité 5', un triplet initie la traduction : il est appelé codon initiateur ou codon start, sa séquence est obligatoirement `ATG` dans le code génétique universel,
– à l'extrémité 3', un triplet déclenche l'arrêt de la traduction : il est appelé le codon terminateur ou codon stop, sa séquence est l'une parmi les trois possibles `TAA`, `TAG` et `TGA`.

FIG. 2.9 – Structure d'un gène codant une protéine.



La séquence d'ARN transcrite est appelée ARN messager ou ARNm car elle sert d'intermédiaire pour l'expression de l'information génétique.

Chez les eucaryotes, certains gènes ont une structure fragmentée qui se retrouve sur l'ARN issu de la transcription. Celui-ci n'est pas directement utilisable pour la traduction et doit subir le mécanisme de l'épissage ("splicing" en anglais).

Épissage Les gènes fragmentés, ou en mosaïque, sont composés d'une suite alternée de régions codantes et de régions non codantes. Seule l'information génétique des séquences codantes est traduite en protéines. Les régions codantes qui quittent le noyau sont appelées les exons[4], et les régions qui restent dans le noyau les introns[5].

Le gène morcelé (cf. fig. 2.10) est transcrit en un ARN pré-messager, ou ARN transcrit primaire, de même composition en exons et introns. Le complexe enzymatique de l'épissage, ou spliceosome, composé de protéines et d'ARNsn, va alors exciser, ou épisser, les introns et ligaturer les exons en formant alors l'ARNm, destiné à la traduction en protéine.

Le splicesome reconnaît les introns grâce à trois séquences nucléotidiques particulières[6] de l'intron et nécessaires à son épissage (cf. fig. 2.11) :

– le site donneur (cf. fig 2.11, 5'SS, pour "splicing site") : séquence située à l'extrémité 5' de l'intron et de longueur 6 nt,

[4]Le terme *exon* provient de *expressed*, exprimé, et du suffixe *-on* repris dans les mots *électron*, *photon*, *etc.*
[5]Le terme *intron* provient de *intragenic*, à l'intérieur du gène, et du suffixe *-on*.
[6]La longueur indiquée des séquences est spécifique des levures.

FIG. 2.10 – Principe de l'épissage d'un ARN pré-messager.
(source : [Cooper, 2000])

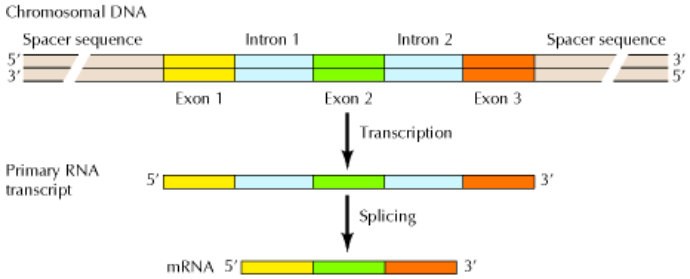
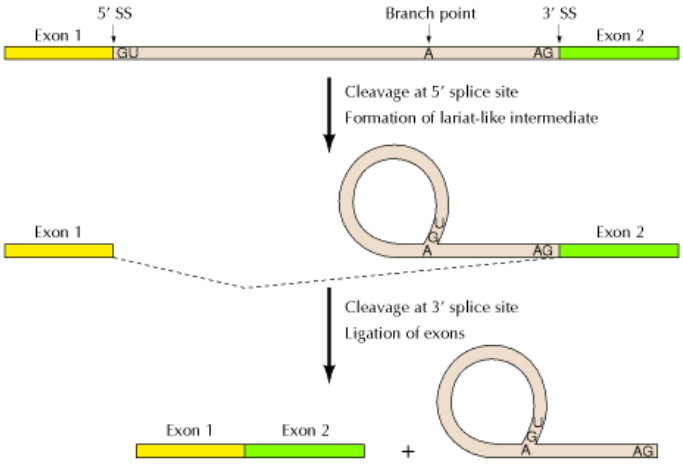


FIG. 2.11 – Épissage d'un ARN pré-messager.
(source : [Cooper, 2000])



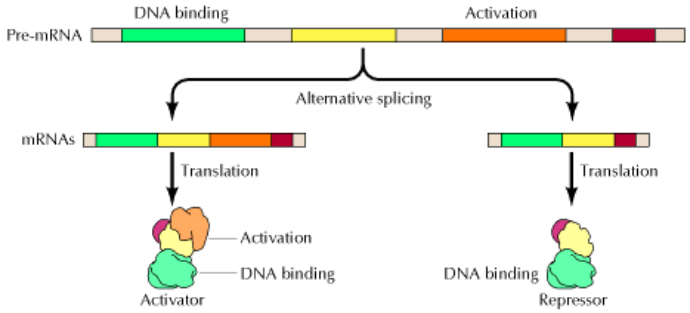
– le point de branchement (cf. fig 2.11, "branch point") : séquence située dans l'intron et de longueur 7 nt, plus proche du site accepteur que du site donneur et
– le site accepteur : séquence située à l'extrémité 3' de l'intron et de longueur 3 nt.

Ces séquences ont, chacune, un motif bien conservé le long de l'arbre phylogénique car ils sont soumis à une forte pression de sélection. Ils permettent, en effet, au spliceosome de s'y fixer puis d'effectuer l'épissage des introns.

Épissage alternatif Des observations chez divers organismes (homme, drosophile, levure...) ont mis en évidence une autre forme d'épissage appelée l'épissage alternatif : un même gène va conduire à des protéines différentes (cf. fig. 2.12).

En fait, le spliceosome reconnaît les signaux d'épissage (des séquences nucléotidiques) pour effectuer l'excision de l'intron. Or ces signaux sont plus ou moins marqués selon leur composition en bases mais aussi selon la présence d'éléments régulateurs de la transcription. Selon le cas, le spliceosome fera ou non l'épissage de tel ou tel intron.

FIG. 2.12 – Épissage alternatif d'un pré-messager.
(source : [Cooper, 2000])



Les protéines finales pourront avoir des séquences distinctes.

L'épissage alternatif est fonction du tissu cellulaire, du stade de développement, des conditions de vie dans un milieu. Une fois épissé, l'ARNm sera dirigé à l'extérieur du noyau pour y être traduit en protéine.

2.3.3 Traduction de l'ARN en protéine

La traduction de l'ARNm en protéine a lieu dans le cytoplasme. Ce processus traduit la séquence de l'ARNm (alphabet de 4 lettres) en une séquence d'acides aminés (alphabet de 20 lettres) selon le code génétique.

La traduction est réalisée par le ribosome (cf. fig. 2.15), complexe nucléoprotéique composé de deux sous-unités, chacune composée de protéines spécifiques dites ribosomiques, et de molécules d'ARN ribosomiques ou ARNr. La petite sous-unité lit l'ARNm, tandis que la grosse sous-unité synthétise la protéine.

Le code génétique (cf. fig. 2.13) établit la correspondance entre un enchaînement de trois nucléotides, aussi désigné par codon ou triplet de nucléotides et un acide aminé. Il y a donc 64 codons possibles (3 positions pour chacun des 4 nucléotides) pour 20 acides aminés : les acides aminés sont ainsi codés par plusieurs codons différents (de 1 à 6), appelés codons synonymes. Le code génétique est qualifié de dégénéré. Par exemple, l'acide aminé méthionine qui est aussi le codon start, est codé par un seul triplet `AUG`, l'acide aminé glutamine est codé par les triplets `GAA` et `GAG`. Trois codons sont non traduits : ils ne codent aucun acide aminé : ce sont les codons terminateurs de la transcription ou codon stop `UAA`, `UAG` et `UGA`.

La longueur de la séquence codante de l'ARNm doit donc être un multiple de 3. Le codon start impose le cadre de lecture dans lequel la traduction est réalisée. Au niveau d'une molécule d'ADN, il existe ainsi 6 cadres de lecture potentiels (cf. fig. 2.14).

Le ribosome synthétise la chaîne protéique à partir des acides aminés amenés, chacun, par une molécule d'ARN de transfert ou ARNt (cf. fig. 2.15). Chaque ARNt possède un triplet nucléotidique particulier déterminant l'acide aminé qu'il porte. Ce triplet de l'ARNt est appelé anti-codon car il correspond à la séquence complémentaire du codon de l'ARNm. Le ribosome, chargé de l'ARNt de la méthionine, se fixe sur l'extrémité 5' de l'ARNm puis se déplace le

FIG. 2.13 – Code génétique universel.
(source : Ijsbrand Kramer et Gérard Tramu)

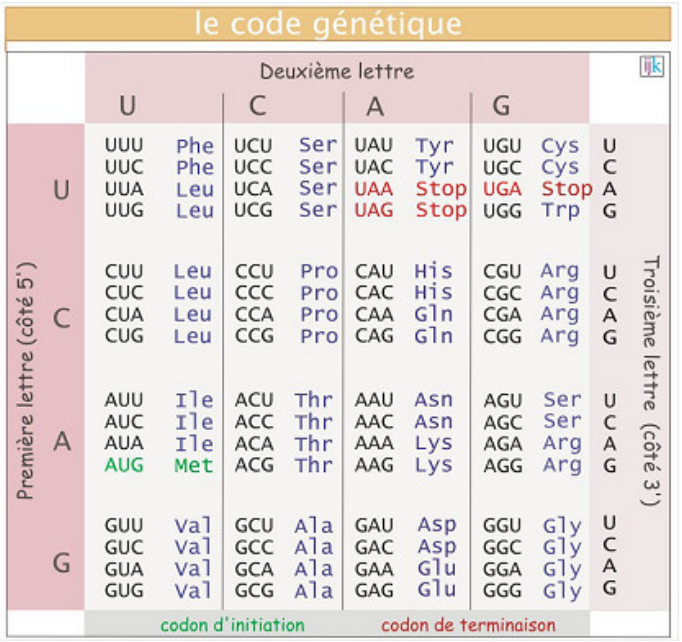
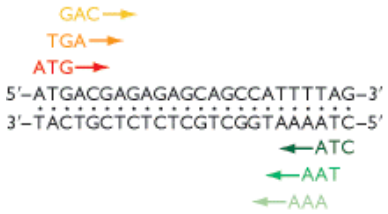


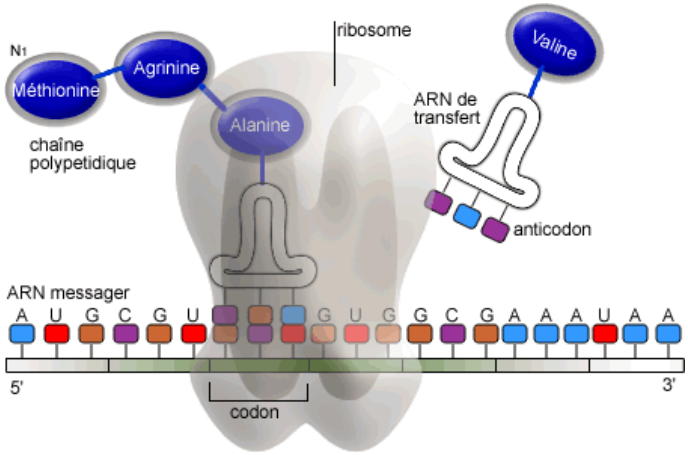
FIG. 2.14 – Les six cadres de lecture de l'ADN.

La molécule d'ADN double brin a six cadres de lecture potentiels. Les deux brins se lisent selon la direction 5' vers 3'. Chaque brin possède trois cadres de lecture, dépendant du nucléotide choisi comme point de départ (source : [Brown, 2002]).



long de la séquence. Dès qu'il rencontre le codon start `AUG` (cf. fig. 2.15, à l'extrémité gauche de l'ARNm), la traduction peut commencer : le ribosome a trouvé le cadre de lecture correct qui lui permet de lire l'ARNm, codon après codon. La partie de l'ARN non traduite en amont du codon start est appelée le 5'UTR (non représentée sur cette figure). Au niveau de chaque codon, l'ARNt ayant l'anti-codon correspondant et chargé de son acide aminé, se fixe sur ce codon (cf. fig. 2.15, fixation de l'ARNt au niveau du codon de l'ARNm). Le ribosome peut alors lier l'acide aminé à la chaîne polypeptidique en cours de synthèse puis passe au codon suivant. Lorsque le ribosome rencontre un codon stop (cf. fig. 2.15, codon `UAA` à l'extrémité 3' de l'ARNm), le ribosome se détache de la protéine et de l'ARNm. La partie de l'ARN en

FIG. 2.15 – Traduction de l'ARNm en protéine.
(source : Guillaume Baukiau)



aval du codon stop est appelée le 3'UTR (non représentée sur cette figure).

En général, un ARNm est lu par plusieurs ribosomes les uns à la suite des autres, avant sa dégradation, amplifiant ainsi le nombre de protéines synthétisées.

La protéine subit ensuite des modifications post-traductionnelles (ou PTM, "Post-Translational Modification"), plus ou moins nombreuses selon la protéine :

- modification de la structure primaire : clivage de la méthionine initiatrice, clivage de sa chaîne,
- modification de la chaîne latérale : modification chimique de l'acide aminé, ajout d'un groupe fonctionnel (glucose, lipide, phosphate...) sur l'acide aminé, ajout par liaison covalente ou non de molécule organique (par exemple l'hème), de molécule nucléique, d'atome métallique...

Les PTM permettent aussi à la protéine d'acquérir ses conformations 2D, 3D et 4D et d'être ainsi entièrement fonctionnelle pour la cellule. Ces modifications post-traductionnelles entraînent un changement de la masse moléculaire de la protéine, parfois très important par rapport à sa masse en amino-acides.

Comme le montrent les mécanismes de la réplication, la transcription et la traduction, les protéines et molécules d'ARN interagissent entre elles afin d'accomplir leur fonction et d'assurer les processus biologiques.

2.4 Interactions moléculaires

La cellule constitue un système complexe où les molécules, protéiques et nucléiques, interagissent entre elles physiquement afin de réaliser et contrôler le fonctionnement cellulaire. Ainsi les interactions moléculaires sont de natures diverses : ADN-ADN, ADN-ARN, ADN-protéine, ARN-protéine et protéine-protéine.

L'ensemble des interactions présentes dans une cellule est appelé interactome ou complexome, en référence aux complexes moléculaires formés par ces interactions.

La formation de complexes est nécessaire à la réalisation de processus biologiques vitaux pour la cellule. Par exemple, le ribosome, qui effectue la traduction de l'ARNm en protéine, est fonctionnel quand ses deux sous-unités sont assemblées. Chacune de ces sous-unités est elle-même un complexe multimoléculaire. La grande sous-unité est constituée de 3 molécules d'ARNr et 49 protéines, la petite sous-unité est constituée d'une molécule d'ARNr et 33 protéines.

De plus, les complexes jouent un rôle important dans le contrôle des processus biologiques dans la cellule. Par exemple, dans la régulation du cycle cellulaire, le passage d'une phase du cycle à la suivante dépend de l'état complexé ou non des protéines impliquées.

Dans le cadre de notre travail, nous présentons ici uniquement la diversité des interactions protéine-protéine et leurs intérêts en terme de spécifité et d'évolution.

2.4.1 Diversité des interactions protéine-protéine (IPP)

D'après Nooren et Thornton [Nooren and Thornton, 2003], les interactions protéine-protéine ou IPP présentent une diversité structurale, fonctionnelle et temporelle.

2.4.1.1 Diversité structurale

Les protéines qui s'assemblent entre elles forment un complexe multi-oligomérique. Il existe différentes sortes de complexes selon la diversité des protéines qui les composent :

- un complexe est homo-oligomérique ou homomultimérique quand il se compose de plusieurs protéines identiques ; par exemple 2 protéines identiques forment un homodimère,
- un complexe est hétéro-oligomérique ou multimérique quand il se compose de plusieurs protéines distinctes ; par exemple 3 protéines différentes forment un hétérotrimère.

2.4.1.2 Diversité fonctionnelle

Les interactions entre protéines peuvent être obligatoires ou non.

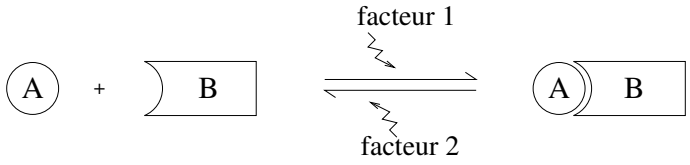
Les composants d'un complexe obligatoire sont instables lorqu'ils sont indépendants dans la cellule. Dans ce cas, les protéines sont fonctionnelles seulement lorsqu'elles sont complexées. Par exemple,le ribosome est actif quand ses deux sous-unités interagissent. De même, les canaux membranaires, qui permettent le passage d'éléments, sont fonctionnels uniquement quand leurs sous-unités sont complexées et localisées dans la membrane.

Dans le cas d'un complexe non obligatoire, les protéines sont stables et fonctionnelles, qu'elles soient complexées ou non. Par exemple, un récepteur et son ligant sont stables et actifs avant même que le ligant ne se lie à son récepteur. Les complexes qui permettent la transduction d'un signal intracellulaire sont également non obligatoires.

2.4.1.3 Diversité temporelle

La durée de vie d'un complexe caractérise celui-ci et peut déterminer sa fonction.

FIG. 2.16 – Équilibre dynamique pour la formation des compexes protéiques.
Les protéines A et B s'associent ou se dissocient selon la variation de facteurs qui peuvent être le pH, une molécule énergétique d'ATP ou GTP, ou même la variation de la concentration des protéines elles-même.



Une interaction permanente est très stable. Les complexes obligatoires sont généralement permanents, car ils ont besoin de cette interaction pour exister.

Une interaction transitoire fait que le complexe existe sous la forme associée et dissociée dans la cellule. Le passage d'un état à un autre peut être lié aux concentrations des composants eux-mêmes : la concentration relative des protéines indépendantes franchit un seuil qui provoque l'agrégation des protéines et inversement. L'interaction transitoire est alors de faible intensité. Le changement d'état peut aussi être lié à la présence d'un facteur externe déclencheur (changement de pH, molécule énergétique d'ATP ou GTP...) (cf. fig. 2.16). L'interaction transitoire est alors de forte intensité. Les complexes non obligatoires peuvent être permanents ou transitoires.

En fait, la formation d'un complexe et sa stabilité dépendent principalement des conditions physiologiques de vie de la cellule et de son environnement. Certains complexes transitoires en conditions normales, deviendront permanents en conditions de stress (stress osmotique, changement de température...). Si les données structurales ou dynamiques au sujet de la formation de complexes restent inconnues, la localisation et la fonction des protéines du complexe permettent de proposer un rôle biologique pour le complexe. Par exemple, les interactions intervenant dans la transduction du signal doivent être transitoires afin de ne pas stimuler la cellule en permanence.

2.4.2 Spécificité des IPP

In vivo, les protéines interagissent entre elles selon leur degré d'affinité et leur localisation (spatiale et temporelle) dans la cellule. Le degré d'affinité repose sur la complémentarité de leur site de liaison, comme une clef dans une serrure. Le site de liaison fait intervenir un ou plusieurs domaines de la protéine. Cette complémentarité dépend de la structure 3D et des propriétés physico-chimiques des protéines mises en jeu. Les facteurs externes tels que les molécules d'ATP, les ions Ca^{2+}, provoquent sur certaines protéines un changement de ces propriétés et de leur conformation spatiale. Par exemple, les interactions intervenant dans la transduction du signal requièrent souvent la présence de ces facteurs, eux-mêmes soumis à des mécanismes de régulation.

La complémentarité doit être parfaite pour permettre une interaction forte et stable. Certaines protéines présentent des sites de liaison peu spécifiques. Les différents partenaires sont alors en compétition pour la fixation sur le site : leur abondance relative dans les différents compartiments cellulaires détermine le type du complexe formé dans ces endroits. Les interactions spécifiques permettent à des partenaires qui ne sont pas co-localisés au début (par exemple un récepteur membranaire et son ligant) de se "trouver".

Les protéines paralogues (protéines homologues regroupées au sein d'une famille) peuvent interagir avec celles d'une autre famille : toute protéine de la famille A interagit avec toute protéine de la famille B. Cette redondance d'interactions possibles est observée par exemple dans les voies métaboliques et les réseaux de régulation tels que la transduction du signal. La cellule assure ainsi le fonctionnement de la voie métabolique en cas de défaillance fonctionnelle d'une des protéines.

Chaque membre d'une famille protéique peut également avoir un partenaire spécifique. Dans ce cas, l'expression temporelle et spatiale joue un rôle important dans la constitution des complexes entre la protéine homologue concernée et son partenaire.

2.4.3 Évolution des IPP

Les IPP assurent des fonctions vitales pour la cellule : réplication, transcription, traduction, transduction du signal, voies métaboliques, régulation... Ces interactions, physiques, se font au niveau des sites de liaison entre partenaires, dépendants des structures 1D, 2D et 3D des protéines en jeu. Toute modification de la séquence nucléique peut avoir d'importantes répercussions sur l'interaction telles que la dérégulation ou la perte fonctionnelle du complexe. Les séquences nucléiques sont donc soumises à une forte pression de sélection naturelle. Du fait de la dégénérescence du code génétique et d'acides aminés partageant des propriétés structurelles et chimiques voisines, certaines mutations n'entraînent pas de modification de fonctionnalité de la protéine mutée ou font même apparaître de nouvelles fonctionnalités utiles pour la cellule [Hoeffken et al., 1988].

Des études d'interaction protéine-protéine à grande échelle sont menées depuis une dizaine d'années chez les organismes modèles tels que la levure, la drosophile, le nématode, dès que leur génome a été connu. Ces études servent de base pour la construction des réseaux d'interaction protéiques et la prédiction des interactions protéine-protéine.

L'étude des complexes protéiques entre diverses espèces, proches d'un point de vue phylogénique, permet de mettre en évidence les différences fonctionnelles et l'adaptabilité des espèces au cours de l'évolution.

2.4.4 Techniques d'étude des IPP

À l'heure actuelle, trois techniques permettent d'étudier les interactions protéine-protéine à grande échelle : le double hybride, la purification en tandem et l'électrophorèse bidimensionnelle en gel bleu natif et SDS.

2.4.4.1 Double hybride

La technique du double hybride [Fields and Song, 1989] (cf. fig. 2.17) repose sur le concept de la proie et de l'appât appliqué dans le contexte de la transcription de gène.

Comme le montre l'étape A de la fig. 2.17, la transcription d'un gène commence si un facteur de transcription se fixe sur une séquence de l'ADN particulière, situé en amont du gène (représentée par le bloc vert sur la fig. 2.17). Le facteur de transcription est une protéine ayant deux domaines distincts :

- un domaine de fixation à l'ADN (DB, "DNA binding", représenté en rose sur la fig. 2.17),
- un domaine d'activation de la transcription (TA, "transcription activator", représenté en jaune sur la fig. 2.17).

Quand une interaction entre deux protéines P_1 et P_2 (représentées respectivement en mauve et en rouge sur la fig. 2.17) veut être testée, la séquence du facteur de transcription est modifiée de telle sorte que :

- la protéine P_1 est exprimée avec le domaine DB : cette protéine joue le rôle de l'appât (cf. fig. 2.17, étape B),
- la protéine P_2 est exprimée avec le domaine TA : cette protéine joue le rôle de la proie (cf. fig. 2.17, étape C).

S'il existe une interaction entre P_1 et P_2, alors le facteur de transcription sera complet et activera la transcription du gène (cf. fig. 2.17, étape D), lequel sera traduit en protéine. Le facteur de transcription choisi est spécifique du gène rapporteur, dont l'activation de la transcription sera facilement décelable. Ce système est appliqué dans une cellule hôte, la levure en général.

L'avantage de cette technique est de déceler des interactions de faible affinité *in vivo*. Mais cette technique présente plusieurs inconvénients. D'abord, elle donne de nombreux faux positifs dans le cas où la proie est peu ou pas assez spécifique (la proie est dite "collante"). Pour gagner en spécificité, la proie et l'appât sont alors réduits à des domaines protéiques et non plus des protéines entières. Ensuite, le système ne tient pas compte des modifications post-traductionnelles présentes dans l'organisme d'origine qui peuvent différer de celles de la cellule hôte. De plus, une seule protéine cible est testée à la fois, ce qui nécessite une automatisation de l'expérience pour étudier les interactions pour l'ensemble des protéines d'une cellule. Cette technique détermine uniquement les interactions entre deux ou trois protéines (système triple hybride [Warbrick, 1997]) et elle est difficilement réalisable avec des protéines membranaires.

Les interactions mises en évidence sont potentielles car elles ne tiennent pas compte des contraintes d'expression en temps et en lieu, ni des concentrations des protéines en jeu ni des facteurs déclencheurs.

2.4.4.2 Purification en tandem

La technique de purification en tandem ou TAP-tag ("tandem affinity purification by tag") [Rigaut et al., 1999, Puig et al., 2001] permet de purifier les complexes protéiques à partir de la cellule d'intérêt, en conservant ainsi les conditions natives d'expression des protéines. Cette technique (cf. fig. 2.18) repose sur une double purification par des colonnes d'affinité.

FIG. 2.17 – Double hybride.

A : le facteur de transcription, composé du domaine de fixation à l'ADN (en pêche) et du domaine d'activation (en jaune), se fixe sur la séquence promotrice du gène et commence la transcription de ce gène. B : la protéine appât (en rose) dont les interactants veulent être connus, est exprimée avec le domaine de fixation : elle joue le rôle d'appât. C : la protéine proie (en rouge) est exprimée avec le domaine d'activation du facteur de transcription. D : l'interaction entre la protéine proie et la protéine appât permet au facteur de transcription d'être fonctionnel : le gène rapporteur est transcrit (source : www.erudit.org).

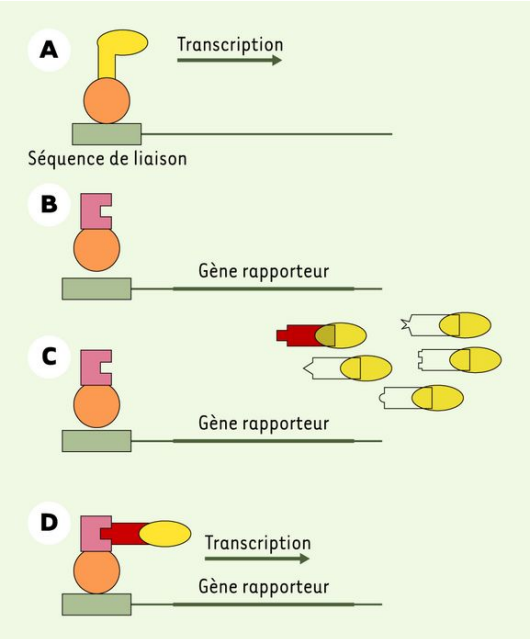
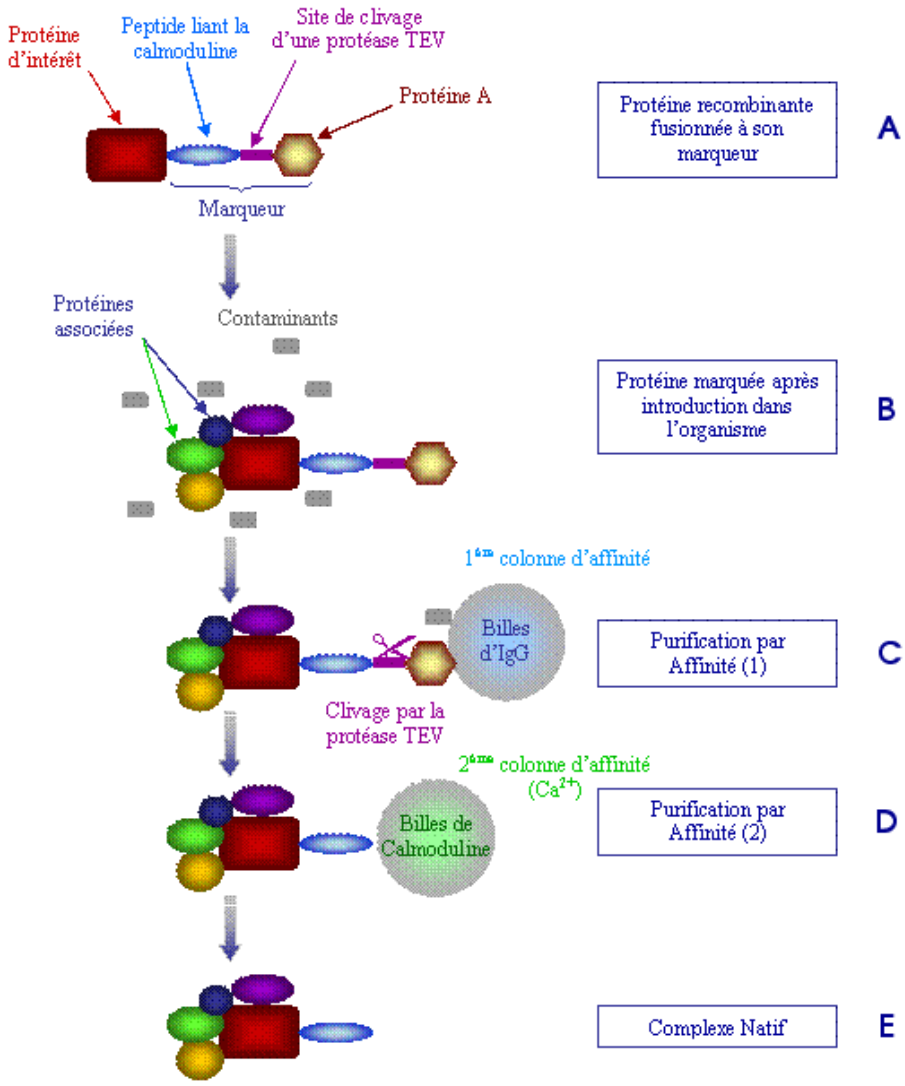


FIG. 2.18 – Purification en tandem
(source : www.erudit.org/)



La double purification se fait par un marqueur composé de trois éléments (cf. fig. 2.18) :

- un peptide liant la calmoduline et couplé à la protéine d'intérêt,
- un site de clivage à une protéase spécifique et
- la protéine A qui possède un site de liaison à l'immunoglobuline G (IgG).

La séquence ADN codant la protéine d'intérêt est fusionnée à celle du marqueur : son expression donne une protéine recombinante qui est exprimée dans la cellule d'origine (cf. fig. 2.18, étape A). Dans la cellule, la protéine interagit avec ses partenaires (cf. fig. 2.18, étape B). Le complexe ainsi formé est purifié à partir d'extraits cellulaires, par "purifica-

tion d'affinité", sur une première colonne d'affinité contenant des billes recouvertes d'IgG (cf. fig. 2.18, étape C). Le complexe, retenu dans la colonne par sa liaison à l'IgG, est élué lorsque la protéase spécifique du site de clivage est ajoutée (cf. fig. 2.18, étape C). Le complexe est ensuite repurifié dans une seconde colonne d'affinité contenant des billes de calmoduline (cf. fig. 2.18, étape D). Le complexe est finalement élué suite à l'ajout d'agents dénaturants (cf. fig. 2.18, étape E). L'éluât est analysé afin de déterminer les partenaires : il est soumis à une électrophorèse de séparation : les protéines vont se séparer selon leur poids moléculaire puis sont identifiées par analyse en spectrométrie de masse.

La technique de l'électrophorèse se base sur le principe de déplacement de molécules chargées électriquement sous l'effet d'un champ électrique. Les protéines migrent dans une matrice solide ou liquide, en fonction de leur charge électrique. La spectrométrie de masse identifie la masse moléculaire d'un composé, permettant de recomposer sa séquence peptidique. Cette séquence est suffisamment longue et caractéristique d'une protéine pour être identifiée par comparaison avec une banque de séquences connues.

Cette technique présente l'avantage de déceler tous les partenaires d'une protéine d'intérêt, à l'échelle de la cellule entière. Cependant, elle a aussi des inconvénients. L'un de ces inconvénients réside au niveau de l'extraction des complexes : les interactions protéine-protéine peu stables peuvent être détruites. De plus, la lyse cellulaire peut aussi induire des interactions artéfactuelles. Un autre inconvénient concerne l'encombrement stérique du marqueur vis-à-vis de la protéine d'intérêt. Celui-ci peut, en effet, gêner certaines protéines dans leur fixation à la protéine d'intérêt.

2.4.4.3 Électrophorèse bi-dimensionnelle en gels bleu natif et SDS

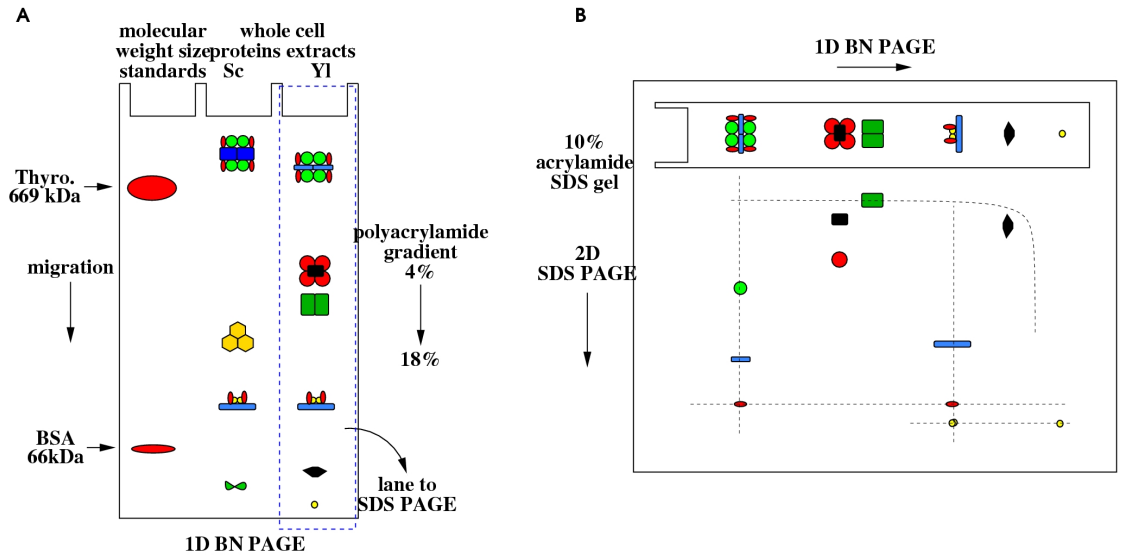
La technique d'électrophorèse bi-dimensionnelle (cf. fig. 2.19) en gels bleu natif et SDS ou BN/SDS PAGE ("Blue Native/SDS PolyAcrylamide Gel Electrophoresis") [Schägger et al., 1996, Camacho-Carvajal et al., 2004] est plus récente que les deux précédentes. Elle se déroule en deux étapes successives :

1. l'électrophorèse en bleu natif (cf. fig. 2.19, étape A) en condition native : les complexes protéiques sont séparés selon leur poids moléculaire et leur forme,
2. l'électrophorèse en SDS (cf. fig. 2.19, étape B) en condition dénaturante due à l'addition du détergent SDS (Sodium Dodecyl Sulfate) : les composants des complexes sont séparés selon leur poids moléculaire.

Un mélange de cellules est lysé en condition native afin de préserver les interactions protéiques des complexes. Cet extrait cellulaire contient ainsi les protéines seules et des complexes protéiques. Ces éléments sont séparés par électrophorèse en condition non dénaturante (cf. fig. 2.19, étape A) sur un gel ayant un gradient linéaire de polyacrylamide. Lors de cette première dimension de séparation, les complexes protéiques et les protéines seules se séparent en fonction de leur masse moléculaire. Une bande de migration du gel est excisée et est dénaturée par SDS. Les IPP liant les protéines complexées sont détruites. La bande est inclue dans un gel pour une seconde électrophorèse en condition dénaturante (cf. fig. 2.19, étape B) : chacune des protéines migre selon sa masse. Les composants des complexes se séparent donc.

FIG. 2.19 – Électrophorèse bi-dimensionnelle en gels bleu natif et SDS.

A. Électrophorèse en bleu natif. Les protéines thyroblobuline (Thyro.) et sérum albumine bovine (BSA) sont des marqueurs de poids moléculaire standards. Deux échantillons d'extraits protéiques obtenus à partir d'un lysat de cellules entières (Sc : *S. cerevisiae*, Yl : *Y. lipolytica*) migrent selon cette électrophorèse de première dimension, en condition native. Les complexes migrent selon leur poids moléculaire. Après migration, une bande est excisée et utilisée pour l'électrophorèse en SDS.
B. Électrophorèse en SDS. L'ajout de détergent SDS provoque la dissociation des protéines complexées. Chaque protéine migre selon son poids moléculaire. Les protéines, révélées par coloration argent, sont excisées et identifiées par spectrométrie de masse.



Chaque composant, localisé par coloration argent, est ensuite excisé du gel et identifié par analyse en spectrométrie de masse.

Lors de la première dimension, le pourcentage de polyacrylamide du gel est proportionnel à la densité du maillage du gel. Ainsi un gradient de 4 à 18% permet de séparer la majorité des complexes d'une cellule sur un même gel [de 40 kDa à 800 kDa]. Le choix du gradient du gel permet de "zoomer" sur une gamme de masse moléculaire.

Cette technique permet de séparer les complexes protéiques à partir d'extraits cellulaires obtenus en conditions naturelles. Mais la phase d'obtention d'extraits cellulaires peut casser les interactions de faible intensité comme en provoquer des fausses. Cependant cette technique a l'avantage de ne nécessiter aucune modification de séquence ADN ni l'utilisation de cellule hôte.

2.4.4.4 Limites quantitatives et qualitatives des techniques expérimentales

Quelle que soit la technique expérimentale utilisée, la totalité des interactions protéiques est difficile à identifier d'un point de vue quantitatif et qualitatif.

Du point de vue quantitatif, même si les techniques sont performantes, il faut un minimum de matériel biologique pour permettre une détection. Or certains complexes intervenant dans des réactions spécifiques sont en trop faible quantité pour atteindre ces seuils de détection. L'identification d'un complexe n'implique pas obligatoirement l'identification de tous ses constituants. Une stoechiométrie (relation quantitative entre les composants du complexe) défavorable pour un composant le rendra d'autant plus difficile à détecter par rapport aux autres.

Du point de vue qualitatif, certaines interactions échappent aux techniques parce qu'elles sont trop fugaces et trop faibles pour pouvoir être observées. De plus, la plupart des produits chimiques utilisés en biologie moléculaire et cellulaire sont nocifs pour la cellule et détruisent les interactions. Les techniques de TAP-tag et du double hybride évitent cela mais elles n'observent pas les interactions en conditions naturelles dans la cellule, entraînant la détection de nombreux faux positifs. La technique BN/SDS PAGE observe les interactions en condition native mais elle n'est pas assez sensible pour pouvoir les détecter toutes.

Il faut combiner les différentes approches et leurs résultats afin d'obtenir une carte des interactions protéine-protéine la plus proche possible de la réalité biologique.

2.5 Conclusion

Les études expérimentales permettent de comprendre les mécanismes physiologiques existant dans une cellule. Le séquençage et l'annotation des génomes ont largement contribué à une meilleure connaissance du fonctionnement de l'organisme. En connaissant les séquences des éléments mis en jeu, certaines expérimentations peuvent cibler l'analyse des gènes et des protéines, ainsi que leurs rôles et leurs relations. D'autres permettent des analyses de plus grande envergure. Le développement des méthodes bio-informatiques permet de traiter et d'analyser le flux continu de données biologiques ainsi générées et d'en extraire de la connaissance qui doit cependant être validée expérimentalement. Grâce à la comparaison entre génomes

proches phylogéniquement, les méthodes *in silico* peuvent, entre autres, prédire les fonctions de nouvelles protéines, extrapoler leurs interactions et leur implication dans les mécanismes cellulaires.

Chapitre 3

État de l'art des stratégies bio-informatiques pour l'annotation génomique

Sommaire

Un moyen d'étude d'un organisme est l'identification des éléments constitutifs de ses chromosomes et de leurs relations fonctionnelles. Ces éléments fonctionnels, inscrits sur la molécule d'ADN génomique, sont les éléments exprimés sous forme de molécules d'ARN et de protéines, et les éléments non exprimés tels que les séquences régulatrices de

l'expression génique. Leurs relations fonctionnelles peuvent être physiques, telles que les interactions protéine-protéine ou ADN-protéine. Elles peuvent aussi porter sur une caractéristique commune, telle que l'appartenance à une famille de gènes ou à un processus métabolique. Les éléments devenus non fonctionnels, tels que les pseudogènes, sont également pris en compte lors de l'annotation car ils témoignent de l'évolution subie par l'organisme. L'identification systématique de tous ces éléments, lors de l'annotation, requiert le séquençage de l'ADN génomique. Puis, les relations entre ces éléments sont obtenues par homologie avec d'autres séquences annotées lors de travaux antérieurs. Bénéficiant de l'existant, l'annotation génomique d'une espèce nouvellement séquencée permet ainsi de disposer d'une quantité d'informations sur ses éléments et une partie de leurs relations, de bonne qualité, sans pour autant engager de nouvelles expérimentations.

Nous précisons, dans ce chapitre, l'état actuel des connaissances relatives à l'annotation génomique et les problèmes auxquels est confronté l'annotateur. Puis nous considérons diverses stratégies bio-informatiques utilisées lors des premiers projets d'annotation génomique et leur degré de satisfaction vis-à-vis des problématiques de rapidité et de qualité présentées en introduction. Nous présentons ensuite les méthodes développées à ce jour pour la vérification de la cohérence de l'annotation. Enfin nous présentons le projet Génolevures, cadre dans lequel s'est déroulé ce travail de thèse.

3.1 Annotation génomique

Le génome d'une cellule eucaryote est le support de l'information héréditaire contenant son programme de fonctionnement. Il contient aussi les informations héritées non fonctionnelles, reliques du processus évolutif subi par cet organisme. L'annotation permet d'obtenir les connaissances sur le fonctionnement cellulaire de l'espèce ainsi que sur les mécanismes hypothétiques de son évolution.

Les levures hémiascomycètes, organismes unicellulaires étudiés ici, contiennent un génome nucléaire et un génome mitochondrial. Ce génome mitochondrial représente une information génétique vitale pour la cellule mais les protéines codées par ce génome n'ont pas d'interactions directes en dehors de la mitochondrie. Ces deux génomes étant de structure identique, leur annotation repose sur les mêmes principes. Pour cette raison, notre discussion se concentre sur l'annotation génomique nucléaire. Néanmoins, l'annotation du génome mitochondrial, de part sa structure d'origine procaryote et l'utilisation d'un code génétique différent, fait l'objet de recherches spécifiques [Cotter et al., 2004, Wyman et al., 2004].

L'annotation d'un génome permet ainsi de déduire ses caractéristiques fonctionnelles et physiologiques fondamentales. Elle résulte de l'enchaînement de deux étapes qui, néanmoins, se recoupent : l'annotation syntaxique puis l'annotation fonctionnelle. Une troisième étape intervient : l'annotation relationnelle, qui relève de l'extraction de connaissance.

3.1.1 Annotation des séquences fonctionnelles et non fonctionnelles

Les éléments importants pour l'annotation sont les élèments fonctionnels : les gènes, exprimés sous la forme de molécules d'ARN et de protéines, et les centromères. Les molécules

d'ARN (ne codant pas de protéine) interviennent au sein de complexes nucléoprotéiques dans divers processus biologiques (traduction, épissage...). Elles ciblent les régions des molécules nucléiques d'intérêt. Le centromère est la zone, unique, du chromosome qui assure une bonne séparation des chromosomes lors de la division cellulaire.

Les séquences non fonctionnelles doivent être également considérées lors de l'annotation car elles permettent de retracer l'histoire de l'espèce. Ces indices évolutifs sont les pseudogènes et les rétrotransposons. Un pseudogène est un gène devenu non fonctionnel suite à une accumulation de modifications délétères de sa séquence ou de son promoteur, séquence régulatrice de l'expression génique. Plus les modifications subies sont nombreuses, plus il est difficile d'identifier la séquence du gène d'origine : le pseudogène est alors une relique de gène [Lafontaine et al., 2004]. Dans le cadre de l'annotation, pour des raisons techniques (cf. § 3.1.3), seuls les pseudogènes ayant leur séquence codante faiblement mutée seront détectés en tant que tels. Les rétrotransposons sont des séquences ADN capables de se déplacer et de se multiplier de façon autonome dans le génome. Les rétrotransposons[1] forment une partie des éléments transposables. Ils provoquent ainsi une instabilité génétique en s'insérant dans les zones codantes telles que les gènes ou les zones régulatrices de l'expression d'un gène. Les séquences nucléotidiques des rétrotransposons et des centromères sont en partie caractérisées chez les organismes séquencés, permettant ainsi leur recherche lors de l'annotation d'un organisme apparenté.

L'annotation des régions non géniques de l'ADN fait actuellement l'objet de recherches. Jusqu'aux années 1990, cet ADN non codant est qualifié d'ADN "poubelle" par opposition à l'ADN codant, considéré alors comme seule véritable information. Par la suite, les détections de différents motifs tels que des motifs répétés [Nadir et al., 1996], des éléments de régulation [Zuckerkandl, 1997] de l'expression des gènes, lui ont conféré une importance désormais égale à celle de l'ADN codant [Bernardi, 1997]. Cependant, l'annotation des régions non codantes s'avère plus difficile que celle des régions géniques du fait de la difficulté d'identification des éléments intéressants dans ces zones. D'une part, comme leur nom l'indique, ces régions sont non exprimées au niveau ARN et protéique. Il est alors plus difficile de mettre en évidence directement, par méthode expérimentale, les effets engendrés par un changement de ces zones (par modification de la séquence nucléique). D'autre part, les séquences non codantes ont une structure moins évidente que les séquences codant des protéines. En l'état actuel des recherches [Dimitri et al., 2005, Castillo-Davis, 2005], ces séquences ne semblent pas répondre à des lois universelles admises par la communauté scientifique comme c'est le cas pour l'ADN codant. Par exemple, elles n'ont pas de bornes précises telles que le codon initiateur et celui terminateur pour un gène codant une protéine. De plus, une région non codante peut contenir divers éléments dignes d'intérêt, par exemple un ou plusieurs transposons provenant de l'insertion passée de virus et rétrovirus, un pseudogène, une séquence promotrice. En 1992, Brosius et Gould [Brosius and Gould, 1992] présentaient une nouvelle nomenclature pour les

[1]Les rétrotransposons bougent dans le génome selon le principe du "copier-coller" en passant par un intermédiaire à ARN (et non ADN comme pour les transposons qui suivent le principe du "couper-coller"). Cela va à l'inverse du dogme central de la biologie (relation unidirectionnelle : "ADN vers ARN vers protéine") régnant à l'époque de la découverte des rétrotransposons par Barbara MacClintock [McClintock, 1953], dans les années 1950, d'où le préfixe "rétro".

pseudogènes et l'ADN "poubelle". En particulier, ils introduisaient le terme de "potonuon" ou "potogene" pour désigner les séquences nucléiques ayant potentiellement une utilité qui serait découverte dans le futur. Plus récemment, Petsko [Petsko, 2003] proposait d'ailleurs de remplacer "junk" par "funk", pour "functionally unknown". Même lorsque ces éléments non codants sont identifiés et localisés, les séquences restantes seront considérées comme "poubelle", non informatives, jusqu'à ce que de nouvelles découvertes infirment à nouveau ce qualificatif.

L'annotation de ces éléments fonctionnels et non fonctionnels fait appel à l'intégration de diverses sources de données à grande échelle, de nature expérimentale et prédictive (par analyse bio-informatique).

3.1.2 Sources de données

L'annotation de génome résulte de l'utilisation de deux sources de données : celles expérimentales provenant des analyses biologiques, et celles prédictives et théoriques issues d'analyses bio-informatiques à grande échelle.

Les premières sources de données, provenant d'expériences biologiques, décrivent la méthode expérimentale, les résultats et leur(s) interprétation(s), et sont publiées dans la littérature scientifique. Des banques de données bibliographiques répertorient ces publications, telle la banque PubMed[2], permettant ainsi d'accéder, via une recherche par mot-clef, aux références complètes d'un article. L'utilisation de références croisées entre banques de données permet, en suivant les liens, d'accéder à l'article.

Les analyses bio-informatiques, secondes sources de données, consistent à rechercher soit des séquences annotées homologues à une séquence inconnue, soit des motifs connus dans la séquence. L'annotateur doit donc cribler, avec ces séquences et motifs, les banques de données où sont stockées ces séquences annotées et autres motifs, ainsi que les références bibliographiques des expériences biologiques.

De façon générale, une banque de données de séquences regroupe, pour chaque séquence, diverses informations biologiques et physico-chimiques. Les informations les plus courantes sont : l'organisme d'origine, la localisation chromosomique, la séquence nucléotidique et sa taille (et s'il s'agit d'un gène codant une protéine : la séquence protéique vérifiée ou prédite, et son poids moléculaire), son annotation fonctionnelle, les références bibliographiques des expériences les plus intéressantes la concernant, les motifs protéiques pertinents (séquence répétée, domaine fonctionnel), ses partenaires d'interaction.

L'annotateur dispose de deux genres de banques de données : les banques de données généralistes et celles spécialisées. Les banques de données généralistes regroupent le maximum de données possible provenant de tous les organismes vivants étudiés. Les banques de données publiques les plus importantes et les mieux tenues à jour en biologie moléculaire sont le NCBI[3] et UniProt[4] [Apweiler et al., 2004, Apweiler et al., 2007]. Elles offrent l'accès aux séquences biologiques. De plus, leurs responsables développent et mettent à disposition des

[2]PubMed est un service proposé par les organismes américains National Library of Medecine et National Institutes of Health : http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=PubMed.

[3]National Center for Biotechnology Information : http ://www.ncbi.nlm.nih.gov.

[4]Universal Protein Resource : http ://www.ebi.uniprot.org/index.shtml.

outils d'analyse bio-informatiques. Elles permettent également l'accès à d'autres banques de données, généralistes ou spécialisées, grâce à des références croisées. UniProt concentre l'ensemble des séquences protéiques publiées et leurs informations. Le NCBI, quant à lui, collecte la littérature des Sciences de la Vie.

Les banques de données spécialisées mettent à disposition les données relatives à un thème précis. Il peut s'agir de banques de données dédiées à un organisme, telle la banque SGD[5] [Cherry et al., 1998] spécialisée pour *S. cerevisiae*, ou à une branche phylogénique, telle Génolevures[6] [Souciet et al., 2000] spécialiste du phylum des Hémiascomycètes. Certaines se focalisent sur les domaines protéiques, comme c'est le cas pour la banque de données Pfam[7] [Sonnhammer et al., 1997], ou bien sur les groupes de protéines homologues entre espèces, telle COG[8] [Tatusov et al., 2000] proposée par le NCBI. D'autres banques regroupent les données d'interactions protéiques, telle la banque IntAct[9] [Hermjakob et al., 2004b], ou les réseaux de réactions et d'interactions moléculaires telle la banque KEGG PATHWAY[10] [Kanehisa et al., 2006].

L'annotateur doit consulter les données les plus pertinentes, lors de chaque étape de l'annotation. Il doit aussi les croiser et les vérifier car ces banques de données peuvent contenir des erreurs [Galperin and Koonin, 1998, Jones et al., 2007] et des informations obsolètes, comme nous l'avons vu en introduction.

3.1.3 Annotation syntaxique

L'annotation syntaxique, ou annotation structurale, identifie les éléments génomiques d'intérêt. Comme nous l'avons vu au § 3.1.1, l'essentiel de ces régions sont les séquences codant des protéines et des molécules d'ARN (ARNt, ARNr, ARNsn...), mais aussi les pseudogènes, les rétrotransposons et les centromères. Certaines séquences répétées et séquences régulatrices de l'expression des gènes peuvent également être détectées grâce à leurs motifs structuraux particuliers. La détection de ces régions d'intérêt se fait à l'aide de prédictions *ab initio* et par comparaison de séquences ou de motifs. La recherche de ces régions se fait dans les trois phases de lecture de chacun des deux brins de la molécule d'ADN. Les prédictions peuvent donc être chevauchantes.

3.1.3.1 Prédictions *ab initio*

La prédiction *ab initio* s'applique aux gènes codant une protéine car ils possèdent une structure précise : un codon start et un codon stop délimitent une séquence en phase de lecture. Cette prédiction peut se baser sur le potentiel codant d'une séquence ou bien la recherche de signaux informatifs d'un gène tels qu'un cadre de lecture ouvert borné par les codons initiateur et terminateur. L'existence chez les eucaryotes d'un cas particulier complique

[5] *Saccharomyces* Genome Database : http ://www.yeastgenome.org.
[6] Génolevures : http ://cbi.labri.fr/Genolevures.
[7] Protein family : http ://www.sanger.ac.uk/Software/Pfam.
[8] Clusters of Orthologous Groups of proteins : http ://www.ncbi.nlm.nih.gov/GOC.
[9] IntAct : http ://www.ebi.ac.uk/intact/site/index.jsf.
[10] KEGG PATHWAY Database : http ://www.genome.ad.jp/kegg/pathway.html.

l'identification de gènes entiers : il s'agit de l'existence de gènes dits fragmentés, composés d'exons et d'intron(s).

D'après la définition d'un gène (cf. page 18), il faudrait également détecter les 5'UTR et 3'UTR. Or ceux-ci ne peuvent être détectés qu'expérimentalement, par analyse du transcriptome. De ce fait, nous ne les considérons pas ici. À ce jour, pour les levures, seuls les UTR de *S. cerevisiae* ont été identifiés à partir d'analyses à grande échelle [David et al., 2006].

Potentiel codant d'une séquence Le potentiel codant d'une CDS peut également être pris en considération en se servant du biais de l'usage des codons [Grantham et al., 1980] propre à chaque espèce. L'étude du code génétique a mis en évidence qu'un même acide aminé peut être codé par plusieurs codons synonymes. Or des études statistiques réalisées sur les différents organismes séquencés jusqu'alors, ont révélé que les fréquences d'apparition des différents codons pour un même acide aminé ne sont pas identiques, pour un organisme donné. Ainsi, chez la drosophile [Griffiths et al., 2002], les deux codons synonymes `UGC` et `UGU` (cystéine) représentent respectivement 73% et 27% des cas. De même, pour les codons correspondant à la valine, `GUG` est utilisé à 48%, tandis que les trois autres `GUC`, `GUU` et `GUA` sont utilisés respectivement à 26 %, 18 % et 8 %. Les profils de ce biais d'utilisation des codons reflètent l'abondance relative des ARNt [Bulmer, 1987, Percudani et al., 1997] ainsi que le niveau d'expression des gènes [Jansen et al., 2003].

Le biais d'utilisation des codons peut être calculé selon deux types de méthodes : (i) des méthodes basées sur la distribution statistique [McLachlan et al., 1984] ; (ii) des méthodes utilisant des séquences codantes et non codantes comme référence, appartenant à l'espèce en cours d'annotation ou à une ou plusieurs espèce(s) apparentée(s) [Angellotti et al., 2007]. Communément utilisé, le programme GeneMark [Borodovsky and McIninch, 1993] détermine le potentiel codant d'une séquence ADN à l'aide d'une fenêtre glissante dans les trois phases de lectures possibles pour chacun des deux brins de l'ADN. Pour cela, il utilise des modèles de Markov (cf. ci-après) créés à partir de régions codantes et non codantes spécifiques d'espèces.

Détection d'un gène Contrairement aux gènes procaryotes, les gènes eucaryotes peuvent être composés d'une alternance de séquences codantes et non codantes, respectivement les exons et introns. Détectée initialement chez les adénovirus[11] [Berget et al., 1977, Chow et al., 1977] puis chez les eucaryotes supérieurs [Pellegrini et al., 1977, Glover and Hogness, 1977, Wellauer and Dawid, 1977] (pluricellulaires), cette structure fragmentée fut également mise en évidence chez les eucaryotes inférieurs (unicellulaires) tels que la levure [Gallwitz and Sures, 1980]. Or l'alternance de séquences codantes et non codante(s) complique grandement l'identification d'un gène entier, comme nous allons le voir par la suite.

Le principe de la détection d'une séquence codante a d'abord été appliqué chez les organismes procaryotes, premiers séquencés, avant d'être appliqué chez les eucaryotes. Initialement, la stratégie utilisée pour la détection de gènes chez les premiers eucaryotes séquencés,

[11] P ;A. Sharp et R.J Roberts reçurent le prix Nobel de Médecine en 1993 pour leurs travaux sur les "gènes divisés".

comportait deux phases successives : (i) la détection d'un cadre de lecture ouvert correspondant potentiellement à une séquence codante ; (ii) puis la prédiction d'introns d'après la présence de motifs d'épissage. À présent, la prédiction de gènes intègre la recherche de séquences codantes et d'introns potentiels grâce à diverses méthodes statistiques. Ces méthodes s'appuient sur les réseaux neuronaux, l'analyse discriminante linéaire ou les modèles de Markov cachés (HMM, "Hidden Markov Model"). Seuls ces derniers sont largement utilisés actuellement, en raison de leur caractère adaptatif, leur grande flexibilité et l'interprétation simple des paramètres du modèle, car le paramétrage influence la qualité de l'annotation.

Nous présentons d'abord la stratégie initiale utilisée pour la détection des gènes. Nous présentons par la suite la stratégie basée sur le modèle de Markov caché.

Détection de cadre de lecture ouvert La détection d'un gène consiste à identifier, dans un premier temps, un cadre de lecture potentiel traduisible en une séquence amino-acide, qui soit délimitée par deux codons stop, ou ORF ("Open Reading Frame", cadre ouvert de lecture). La recherche d'ORF est réalisée dans les trois cadres de lecture de chacun des deux brins d'ADN.

L'annotateur paramètre la recherche d'ORF sur la taille minimale d'un ORF à détecter. En effet, plus la taille choisie de l'ORF est petite, plus l'annotateur aura d'ORF à analyser et à annoter. De ce fait, le nombre de faux positifs augmentera en conséquence. Le choix de la longueur minimale de l'ORF pour un organisme inconnu est basée sur des observations faites sur l'organisme de référence le plus proche ou des organismes apparentés. Par exemple, lors de l'annotation des génomes de levures nouvellement séquencés, la longueur minimale d'un ORF est fixée à 60 nt, soit une séquence codant une protéine de 20 aa, en référence à *S. cerevisiae* dont les plus petites protéines identifiées ont une taille de 24 aa (YBR191W-A et YDL247W-A [Blandin et al., 2000]) et 16 aa (YJR151W-A [Kessler et al., 2003]).

Puis, à l'intérieur de cet ORF, la présence d'un codon initiateur en phase avec le cadre de lecture de l'ORF, permettra la prédiction d'une CDS, séquence prédite codant une protéine délimitée par le codon initiateur et terminateur. L'annotateur peut également fixer, selon la même méthodologie, une longueur minimale pour la CDS, réduisant ainsi le nombre de faux positifs. D'après l'analyse du génome de *S. cerevisiae* (6 577 CDS réparties sur les 12,07 Mpb constituant les 16 chromosomes de l'ADN nucléaire[12]), la taille moyenne d'une CDS est d'environ 1,1 kpb, donnant un taux de couverture de la partie codante proche de 60% du génome. Ce taux de couverture génique est extrêmement variable selon les branches phylogéniques (1,8% chez l'homme) ; plus il est faible, plus les recherches des séquences codantes dans un génome sont complexes.

Prédiction d'épissage Chez les eucaryotes tels que les levures, la détection de séquences codantes à partir de méthodes *ab initio* est compliquée par la présence d'une ou plusieurs séquences non codantes, les introns, dans la séquence codante. Les critères définissant alors une CDS (codons start et stop encadrant une séquence traduisible dans sa totalité

[12]D'après la base de données SGD, au 25/08/2007.

en protéine) ne suffisent plus. En effet, les introns n'ayant pas de contrainte de phase de lecture, les exons peuvent se trouver dans des cadres de lecture différents.

De plus, le codon start se situe majoritairement dans le premier exon[13], ou bien à cheval sur le premier et le deuxième exon (*e.g.* YIL111W [Cumsky et al., 1987] , YJL041W [Hurt, 1988], YPR043W [Waskiewicz-Staniorowska et al., 1998] chez *S. cerevisiae*).

La séquence codante issue de l'épissage des introns doit, quant à elle, pouvoir être traduite en protéine.

Les introns sont détectés grâce à leurs trois motifs caractéristiques (cf. p.23) : le site donneur, le site accepteur et le point de branchement. Ces sites étant nécessaires pour l'interaction entre le spliceosome et la séquence intronique à épisser, leurs motifs sont bien conservés chez l'ensemble des espèces appartenant à une même branche phylogénique, chez les vertébrés [Yeo et al., 2004] ou chez les levures [Bon et al., 2003, Neuvéglise, 2005]. Mais les différentes séquences possibles pour ces motifs, et l'absence de contrainte sur la séquence entre ces sites (en longueur et en contenu), sont suffisantes pour produire un grand nombre de combinaisons satisfaisant ces conditions, générant potentiellement autant de faux positifs que l'annotateur doit éliminer. Comme pour la prédiction d'ORF et de CDS, l'annotateur peut rajouter des contraintes arbitraires sur les motifs et les distances entre ces motifs. Ces contraintes sont issues d'observations statistiques réalisées chez d'autres espèces apparentées et permettent de diminuer le nombre d'introns faux positifs.

L'identification de l'ensemble des exons et introns composant un gène, basée sur la propriété d'alternance exon / intron, est complétée par la recherche de séquences homologues (cf. § 3.1.3.2) à l'ADN génomique. Cependant, d'une part, d'après l'observation de la composition en exons et intron(s) de gènes fragmentés de plusieurs levures [Neuvéglise, 2005], les introns se localisent statistiquement en 5' de la CDS : la séquence codante de l'exon 1 est plus petit que les exons suivants. D'autre part, les algorithmes d'alignement utilisent une fenêtre glissante et des paramètres statistiques. L'exon 1 peut donc être "noyé" dans la fenêtre glissante, ne satisfaisant pas les seuils statistiques requis par l'algorithme pour considérer la présence d'un alignement significatif. En conséquence, un exon 1 de petite longueur peut ne pas être détecté lors d'un alignement avec une séquence homologue. La prise en compte du 5'UTR, que ce soit au niveau de la CDS prédite ou des séquences homologues recherchées, améliorerait l'homologie et donc la qualité de l'annotation, car la recherche de l'alignement se ferait sur une séquence plus étendue.

L'existence d'épissage alternatif (cf. p. 25) d'ARN complique davantage la tâche de l'annotateur. En effet, les différents produits du gène peuvent ne pas être dans la même phase de lecture et être donc prédits chevauchants.

Modèles des chaînes de Markov cachés L'analyse des génomes des différents organismes eucaryotes séquencés à ce jour, a révélé l'existence de gènes avec intron(s) dans les différentes branches phylogéniques du domaine eucaryote. La méthode *ab initio* de prédiction de gène doit alors prendre en compte la détection d'introns dès la recherche d'une séquence codante. Cette méthode fait appel à une approche probabiliste de détection de gènes, basée

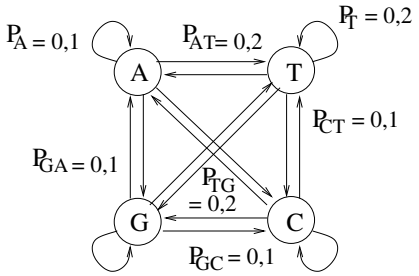
[13]Les 5' UTR n'étant pas détectés, nous ne considérons pas le cas des introns présents dans ces UTRs.

FIG. 3.1 – Ensemble de transitions pour les nucléotides selon une chaîne de Markov.

Chaîne de Markov d'ordre 1 : l'état suivant ne dépend que de l'état précédent. Seules les transitions de la séquence S sont indiquées sur le schéma, les autres ayant des probabilités nulles.

Soit la séquence S:
AATTGCTGATT

P_X probabilité de transition de X vers X
P_{XY} probabilité de transition de X vers Y



sur les modèles de Markov cachés.

Les modèles de Markov cachés étaient déjà utilisés pour la détection des séquences codantes chez les bactéries [Borodovsky et al., 1995] avant d'être utilisés pour la prédiction de gènes complets (exons et introns) chez l'homme grâce au programme GENSCAN [Burge and Karlin, 1997]. Actuellement, cette méthode est utilisée par divers logiciels [Allen et al., 2006] tels que SLAM [Alexandersson et al., 2003], AUGUSTUS [Stanke, 2003], GlimmerHMM [Majoros et al., 2004], JigSaw [Allen and Salzberg, 2005], GeneZilla (anciennement TIGRscan) [Majoros et al., 2004, Majoros et al., 2005].

Principe du modèle de Markov caché

Dans une séquence ADN, l'ordre de succession des nucléotides n'est pas aléatoire : un modèle peut ainsi définir une probabilité pour laquelle une base dépend des bases précédentes.

Une chaîne de Markov (cf. fig. 3.1) est un ensemble d'états et d'arcs. À chaque état est associé un nucléotide (A, T, G ou C), et à chaque arc est associée une probabilité de transition d'un état vers un autre.

La probabilité qu'une séquence satisfasse un modèle donné, est le produit des probabilités de transitions attachées à chacune des transitions apparaissant dans la séquence.

Une chaîne de Markov se caractérise par son ordre k : l'état suivant dépend des k états précédents. La séquence considérée est alors de longueur $k + 1$ nt.

Le concepteur du modèle fixe les probabilités de transition par une méthode d'apprentissage et un jeu de données de référence, composé de séquences homologues. Une nouvelle séquence est homologue à ces séquences si elle satisfait le modèle ainsi paramétré.

Ce test de satisfaction des contraintes du modèle se complique si la séquence contient le motif recherché mais que celui-ci est borné de séquences différentes et sans rapport avec les séquences définissant le modèles, telles que les introns vis-à-vis des exons. Il faut alors définir autant de modèles que de motifs séquentiels (*e.g.* exon, sites d'épissage, région non codante) distincts. La probabilité de passage d'un modèle à un autre est alors calculée à partir des probabilités de transition d'un état d'un modèle à un état d'un autre modèle, et ce, pour

chacun des états des modèles. Il n'y a plus de correspondance directe entre les nucléotides et les états, car pour un même nucléotide, plusieurs transitions vers des états de modèles différents existent. Il sera alors impossible de savoir quel modèle a conduit à quelle séquence : la succession des transitions d'un état d'un modèle à un autre état est inconnue. Les modèles ont alors des états cachés.

Soit la séquence ADN de référence cachée S composée de régions connues (exon, intron, séquence intergénique...). La séquence cachée $S = (S_1, S_2, \ldots, S_n)$ décrit l'alternance de ces régions (*e.g.* région non codante, exon, intron, exon).

Cette séquence est modélisée par une chaîne de Markov (ici d'ordre 1) : la probabilité a d'emprunter la transition d'une région u vers une région v est égale à la probabilité d'avoir v sachant qu'en amont, il y a u :

$$a(u, v) = P(S_t = v | S_{t-1} = u)$$

Chaque modèle est ainsi caractérisé par l'ensemble de ses probabilités de transition d'une région vers une autre.

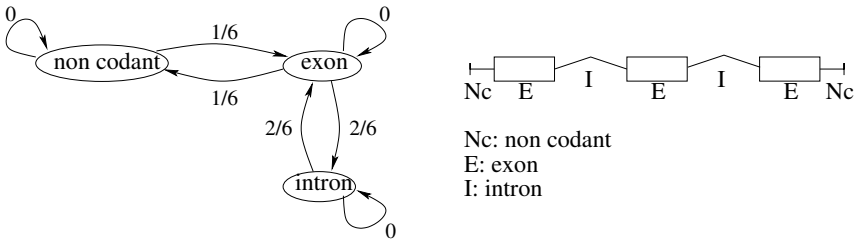
La séquence observée $X = (X_1, X_2, \ldots, X_n)$ correspond à la séquence ADN dont la composition en régions doit être déterminée. Elle est modélisée d'après les conditions imposées par la séquence cachée S. La probabilité b de passer d'une région w à une autre z sur la séquence ADN observée est conditionnée par ce qui est connu sur la séquence cachée S en position t et en tenant compte des k états précédents :

$$b_u(w, z) = P(X_t = w | S_t = u, X_{t-k}^{t-1} = w)$$

La figure 3.2 représente une séquence cachée avec sa modélisation en chaîne de Markov cachée.

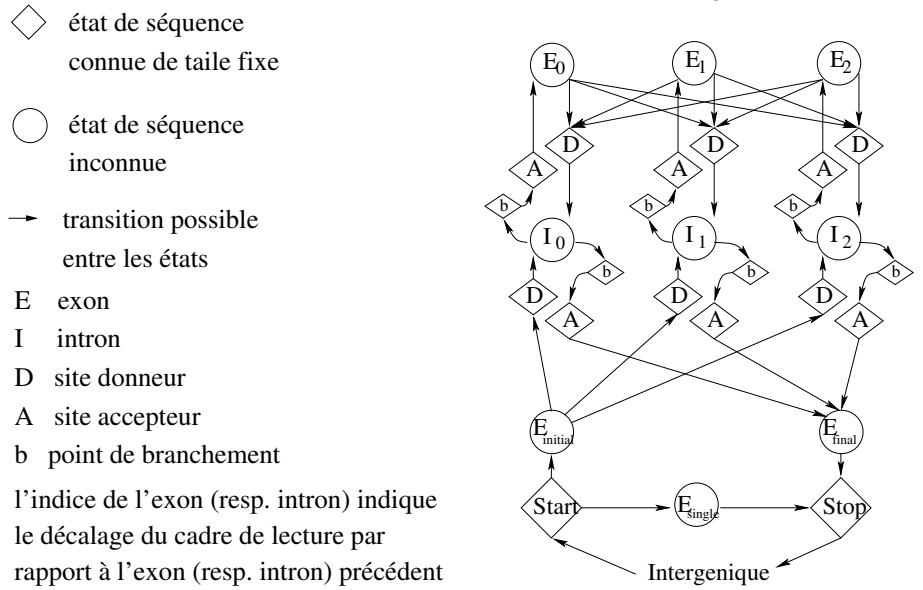
FIG. 3.2 – Modèle de Markov caché.

La séquence cachée ci-dessous définit le modèle de Markov caché composé de trois états : non codant (Nc), exon (E) et intron (I). Les probabilités de transition indiquées sont définies d'après la séquence cachée.



En pratique, un jeu de données de séquences connues (*e.g.* exons, introns, séquences intergéniques...) et apparentées à la séquence à annoter, sert d'apprentissage pour créer les différents modèles. La décomposition en états peut encore être complexifiée par la prise en compte des motifs introniques en 5', 3' et point de branchement. Ces motifs ont une taille fixe

FIG. 3.3 – Modèle de Markov caché utilisé par GeneZilla [Allen et al., 2006].



et peuvent être modélisés par un seul état composé d'autant de bases. Le modèle de Markov devient alors un modèle généralisé.

La séquence à annoter est ensuite modélisée selon ces modèles. La figure 3.3 représente le principe du HMM sur lequel sont basés les logiciels de prédiction de gènes.

Ce modèle markovien ne peut identifier une région de taille trop petite, car la probabilité associée à une telle région sera trop faible. De plus, un gène dont le codon start est à cheval sur le premier exon et le deuxième exon, comme nous l'avons vu p. 45 (§ 3.1.3.1), ne sera pas détecté car ce cas n'est pas pris en compte par ce modèle. En effet, cette prise en compte entraînerait une perte de précision du modèle. Actuellement, de nouvelles approches [Saeys et al., 2007] se penchent sur l'identification des petites séquences codantes, en particulier les courts exons, dans les principaux règnes (vertébrés, plantes, champignons et protistes).

3.1.3.2 Prédictions par recherche de séquences homologues et de motifs

La prédiction de séquences d'intérêt biologique par des méthodes *ab initio* peut se compléter par la recherche de motifs particuliers et par comparaison de séquences. Cette comparaison repose sur le principe de l'homologie de séquences. L'homologie de séquences est une prédiction basée sur l'observation d'une ressemblance entre séquences par comparaison de ces séquences. La similitude de séquences s'appuie sur une relation phylogénique entre ces séquences homologues pour vérifier et confirmer cette ressemblance.

Cette recherche de séquences homologues et de motifs consiste ainsi en l'identification des éléments mentionnés au § 3.1.1 dont la structure n'est pas aussi bien connue que celle des gènes codant les protéines, tels que les rétrotransposons, les gènes codant les ARN et les

centromères. L'annotateur peut aussi disposer de séquences mises en évidence lors d'études expérimentales antérieures à l'annotation en cours. Il peut également rechercher des séquences homologues à celles en cours d'annotation dans des banques de séquences provenant d'espèces apparentées. Cette recherche par homologie de séquences est étroitement liée à l'annotation fonctionnelle. Elle permet de gagner du temps en attribuant une fonction à la séquence prédite par inférence. Nous détaillons cette partie dans le § 3.1.4 ci-après.

L'ensemble de ces analyses permet de proposer à l'annotateur plusieurs éléments codant une molécule d'ARN ou une protéine. Ces propositions sont appelées des modèles de gène. Ensuite l'annotateur juge la validité d'un modèle de gène codant une protéine selon plusieurs critères, principalement : la longueur de la séquence codante, la présence de séquences homologues dans d'autres organismes, un fort potentiel codant, et éventuellement son expérience. De plus, sauf cas exceptionnel, les gènes ne se chevauchent pas chez les eucaryotes. Cette contrainte biologique permet ainsi à l'annotateur de valider un seul ORF parmi ceux chevauchant une même région d'ADN et présents sur les trois phases de lecture des brins sens et anti-sens. Cependant le chevauchement de séquences biologiques a été mis en évidence par expérimentation, par exemple chez l'homme [Coriton et al., 2000]. En sus, le problème de l'épissage alternatif doit toujours être considéré.

Une fois les éléments codants localisés et identifiés, l'annotateur rédige un commentaire fonctionnel d'après des analyses de recherche d'homologie avec d'autres éléments. Ce commentaire peut être rédigé en se servant d'un vocabulaire contrôlé afin d'avoir une annotation homogène, gage de qualité.

3.1.4 Annotation fonctionnelle

L'annotation fonctionnelle a pour but de prédire les fonctions des produits des gènes identifiés lors de l'annotation syntaxique. Un commentaire décrivant cette fonction potentielle est ainsi ajouté, *e.g.* le gène prédit DEHA0C12331g a pour annotation foncionnelle "`highly similar to CA4862|CaEFB1 Candida albicans CaEFB1 translation elongation factor eEF1beta, start by similarity`" : DEHA0C12331g pourrait avoir une fonction similaire à celle d'un facteur d'élongation de la traduction. Les éléments, pris individuellement et présentant les caractéristiques fonctionnelles similaires, pourront alors être regroupés en familles : familles de gènes, de protéines, de séquences régulatrices... Si des études expérimentales, antérieures au séquençage complet du génome, concernent ce génome et sont disponibles, l'annotateur attribue de façon certaine des informations fonctionnelles aux éléments concernés. À l'exception des organismes modèles, l'annotateur dispose de peu d'informations pour les organismes nouvellement séquencés. Il doit alors interroger les banques de données, à la recherche de séquences ou de motifs homologues à la séquence inconnue afin de lui assigner, par inférence, leur(s) fonction(s).

Définition de l'inférence entre gènes Soit g_{ESP} un gène connu d'une espèce ESP ayant la propriété *propriete*, et g_X le gène prédit de l'espèce X en cours d'annotation et homologue à g_{ESP} :

Définition 3.1.1 *Soit un couple* (g_{ESP}, g_X) *vérifiant* $homologie(g_{ESP}, g_X)$, *alors* $propriete(g_{ESP}) \Rightarrow propriete(g_X)$.

Seules certaines propriétés peuvent être extrapolées par inférence, par exemple la fonction et l'interaction. Ainsi l'annotation fonctionnelle d'un gène prédit peut être induite à partir d'un gène homologue. Elle repose d'une part sur la recherche d'homologie entre la séquence inconnue considérée et un ensemble de séquences annotées, et d'autre part sur la recherche de motifs particuliers au sein de cette séquence.

3.1.4.1 Recherche de séquences homologues

La comparaison directe de la CDS avec un ensemble de séquences annotées permet de sélectionner les séquences similaires pour lesquelles des informations fonctionnelles sont disponibles et d'en déduire des informations pour cette CDS.

La comparaison d'une séquence (ADN ou protéine) avec celles d'une banque de données (ADN ou protéine) donne un score statistiquement significatif qui permet de trier les homologies de séquences selon leur probabilité d'être observée par hasard (plus ce score est faible, plus la probabilité que l'homologie entre les séquences soit due au hasard est grande). L'algorithme d'alignement local de Smith et Waterman [Smith and Waterman, 1981] est le plus répandu et connu des algorithmes de comparaison de séquences. Il a été implémenté et amélioré dans les logiciels de recherche d'homologie de séquences tels que le logiciel BLAST [Altschul et al., 1990].

La comparaison directe de séquences peut aussi tenir compte de données phylogéniques, afin de cibler la recherche d'homologies de séquences.

3.1.4.2 Recherche de motifs

La recherche de motifs dans une séquence inconnue a le même but que la recherche de séquences homologues : obtenir des informations sur les fonctions et propriétés des gènes et produits de ces gènes. L'analyse porte ici sur la composition en bases (nucléiques ou aminoacides) et la structure (présence de domaines responsables d'une fonction, repliements en deux ou trois dimensions) de ces éléments. Elle se fait de deux façons. La première est la confrontation de comparaison de la séquence requête avec celles d'une banque d'éléments fonctionnels (*e.g.* domaines, motifs, structures 2D/3D...). La seconde est le calcul des propriétés physico-chimiques de la séquence inconnue (*e.g.* une protéine de profil hydrophobe a une forte probabilité d'être localisée dans une membrane).

L'annotation fonctionnelle de la CDS peut être complétée par diverses indications basées sur des ontologies biologiques particulières développées par le consortium Gene Ontology [Ashburner et al., 2000] depuis 1998. Ce standard de fait est utilisé dans les domaines de recherche biologique. Ces ontologies [Consortium, 2001] décrivent les attributs des produits de gènes dans trois domaines distincts de la biologie moléculaire : la fonction moléculaire, le processus métabolique, et le composant cellulaire (la localisation cellulaire). Grâce à l'utilisation de ce vocabulaire structuré et contrôlé, ces ontologies permettent aux biologistes de considérer

comme une seule entité un élément présent dans plusieurs banques de données. Chaque ontologie GO est représentée par un réseau particulier de nœuds, appelé graphe dirigé acyclique (DAG, "Directed Acyclic Graph"). Chaque nœud correspond à un terme de l'ontologie qui peut être le fils d'un ou plusieurs parent(s) selon la relation "`est_un`" ou "`est_une_part_de`".

L'ensemble de ces recherches (homologie ou analyse intrinsèque de la CDS) permet ainsi d'identifier plusieurs informations utiles pour comprendre la fonction du gène et de son produit prédits. Cependant la prédiction devra être validée expérimentalement pour en garantir la qualité. Quoi qu'il en soit, cette stratégie d'attribution de fonctions par homologie présente au moins trois limites principales. Malgré la diversité d'analyses disponibles, la CDS peut ne ressembler à aucune autre séquence identifiée à ce jour, ni contenir d'éléments caractéristiques (domaines, motifs...), ce qui fait de lui un gène orphelin, sans annotation fonctionnelle. De plus, il ne faut pas oublier que l'annotation fonctionnelle procède de l'intégration de données hétérogènes qui doivent être recoupées. Transposer l'annotation fonctionnelle d'une séquence mal annotée sur une nouvelle va propager cette erreur, qui sera, elle-même, utilisée ultérieurement. De même, à partir d'un résultat d'homologie de séquence significatif, il est nécessaire de s'assurer de la présence d'un domaine fonctionnel, par exemple responsable d'une activité biochimique sur la CDS, avant d'attribuer cette fonction à cette CDS.

Dès lors que l'annotateur a commenté l'ensemble des gènes et produits des gènes d'un génome, la majeure partie de l'annotation peut être considérée comme terminée. L'annotation obtenue offre alors un catalogue des éléments codants du génome. Mais il n'y a pas de vision d'ensemble du génome en tant que système fonctionnel. La détermination des relations potentielles entre ces éléments s'effectue au cours de l'annotation dite relationnelle. La prise en compte de cette phase dans l'annotation est apparue lorsque les techniques expérimentales ont permis d'étudier, à grande échelle, les interactions entre éléments d'une cellule [Uetz et al., 2000, Ito et al., 2001, Gavin et al., 2002, Ho et al., 2002]. Historiquement, l'annotation relationnelle est considérée comme à part des deux premières phases d'annotation, car elle constitue déjà l'extraction de connaissances à partir des éléments identifiés d'un génome.

3.1.5 Annotation relationnelle

L'annotation relationnelle, ou contextuelle, fait appel à des informations plus complexes que les informations rattachées aux séquences. Elle détermine les relations susceptibles d'exister entre les éléments prédits et caractérisés auparavant. Ces relations sont de diverses natures :

- homologie : les protéines peuvent être regroupées en familles d'homologues, constituées d'après des analyses informatiques, par exemple les algorithmes de clustering consensus [Nikolski and Sherman, 2007], et des expériences biologiques,
- interaction physique : les éléments interagissent physiquement entre eux : protéine/acides nucléiques, protéine/protéine, acides nucléiques/acides nucléiques,
- implication commune dans un processus biologique : participation à la même voie métabolique, même voie de transport, même réseau de régulation.

Pour ce faire, l'annotation relationnelle nécessite soit la mise en place d'expériences biologiques à grande échelle pour cet organisme, soit l'utilisation de méthodes prédictives par inférence,

à partir d'observations chez d'autres organismes. L'annotation par inférence permet ainsi d'enrichir l'annotation d'un organisme sur lequel peu d'informations expérimentales validées sont disponibles.

Dans le cadre de nos travaux, nous présentons ici seulement l'étude des interactions protéine-protéine (IPP). Les IPP peuvent être considérées comme l'un des points essentiels de l'annotation relationnelle, avec les processus biologiques. En effet, elles permettent de définir plusieurs aspects du fonctionnement de la cellule, tels que les voies métaboliques, la transduction du signal, la régulation des processus cellulaires (*e.g.* cycle cellulaire).

3.1.5.1 Méthodes expérimentales d'étude d'interaction protéine-protéine

Comme nous l'avons vu dans le chapitre 2 (cf. § 2.4.4), trois techniques expérimentales éprouvées permettent l'étude des IPP à grande échelle : la technique du double hybride, la purification en tandem (ou TAP-tag) et l'électrophorèse bi-dimensionnelle en gel bleu natif et SDS. Chacune de ces techniques présente ses avantages et inconvénients (cf. § 2.4.4).

La technique du double hybride et le TAP-TAG sont plus longues à mettre en pratique que l'électrophorèse BN/SDS car elles nécessitent la modification de la séquence nucléotidique des gènes codant les protéines proies et appâts.

La technique du double hybride fait appel à une cellule hôte pour exprimer les protéines proies et appâts, en général *S. cerevisiae*. Dans le cadre du projet Génolevures, cela n'a pas vraiment de conséquence car nous pouvons supposer que les mécanismes de modifications post-traductionnels sont conservés au sein d'une branche phylogénique homogène telle que celle des levures hémiascomycètes. Des travaux ont mis en évidence le manque de sensibilité (proportion de faux positifs détectés) et de spécificité (proportion de faux négatifs) de cette technique du double hybride chez *S. cerevisiae* [Legrain et al., 2001, von Mering et al., 2002, Sprinzak et al., 2003]. Les scientifiques se sont alors intéressés à des méthodes informatiques afin de détecter ces faux positifs [Serebriiskii and Golemis, 2001, Chen et al., 2006].

Le TAP-tag et l'électrophorèse BN/SDS permettent l'étude des IPP directement à partir de la cellule de l'organisme choisi, sans passer par une cellule hôte. Cela constitue un avantage non négligeable quand l'organisme étudié est très éloigné, d'un point de vue phylogénique, de la cellule hôte.

Cependant le TAP-tag, comme le double hybride, ne permet l'identification que des partenaires des interactions ciblées. L'électrophorèse BN/SDS ne fait pas de purification intermédiaire : elle permet l'identification de l'ensemble des protéines de la cellule, dans la limite du seuil de détection de la technique. Ceci représente un avantage pour un projet d'annotation génomique car l'identification de ces protéines, complexées ou non, constitue une validation de la qualité de l'annotation syntaxique et fonctionnelle réalisée en amont.

3.1.5.2 Méthodes prédictives d'interaction protéine-protéine

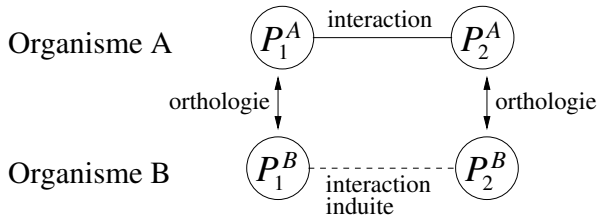
Les réseaux d'interaction protéine-protéine (IPP), ou interactome, peuvent être induits selon diverses méthodes parmi lesquelles l'utilisation de l'orthologie entre les gènes, les événements de fusion de gènes, ou bien le profil phylogénique des protéines. Ces méthodes, utilisées

ensemble [Marcotte et al., 1999, Valencia and Pazos, 2002], permettent d'affiner les prédictions.

Relation d'orthologie entre gènes La méthode la plus simple est l'inférence d'interactions à partir de relations d'orthologie avec un organisme de référence. L'orthologie est la relation d'homologie existante entre deux protéines appartenant chacune à un organisme différent.

Soit un organisme A ayant deux protéines P_1^A et P_2^A qui interagissent entre elles, et un organisme B ayant deux protéines P_1^B et P_2^B orthologues respectivement à P_1^A et P_2^A (nous supposons que chaque gène a au plus un orthologue), une interaction entre P_1^B et P_2^B peut être induite d'après la relation d'orthologie entre ces protéines (cf. fig. 3.4).

FIG. 3.4 – Inférence d'interaction protéine-protéine

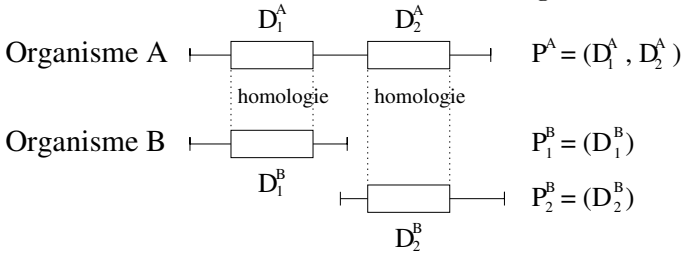


Les membres du réseau d'interaction (ou d'un complexe protéique) de l'organisme B peuvent ainsi être induits sans expérimentation par inférence du réseau de l'organisme apparenté A.

Événement de fusion de gènes L'interaction entre protéines peut être induite à partir de la présence dans au moins deux génomes, des même domaines protéiques qui sont, soit sur une seule protéine, soit sur des protéines distinctes.

Soit un organisme A ayant une protéine P^A composée de deux domaines protéiques D_1 et D_2, et un organisme B ayant deux protéines P_1^B et P_2^B composées respectivement des même domaines D_1 et D_2, une interaction potentielle peut être inférée entre P_1^B et P_2^B (cf. fig. 3.5).

FIG. 3.5 – Événement de fusion de gènes



Cette méthode se base sur la recherche récursive de séquences (ici, recherche des séquences de P^A, P_1^B et P_2^B) et d'alignements multiples entre ces séquences (ici, alignements de P^A avec

P_1^B, de P^A avec P_2^B, mais absence d'alignement, d'après les critères définis par l'annotateur, entre P_1^B et P_2^B) pour détecter ainsi les événements de fusion de domaines protéiques.

Les événements de fusion de gènes ont été observés, entre autres, entre la bactérie *E. coli* et *S. cerevisiae* [Berger et al., 1996, Marcotte et al., 1999]. Par exemple, les séquences homologues des sous-unités Gyr1 et GyrB de la DNA gyrase de la bactérie, sont fusionnées chez la levure, constituant ainsi les deux domaines de la topoisomérase II. Ces événements de fusion de gènes ont été mis en évidence pour des protéines impliquées dans les voies métaboliques, les voies de signalisation, les complexes protéiques.

Profil phylogénique des protéines Cette méthode prédictive [Pellegrini et al., 1999, Vert, 2002] détecte les interactions fonctionnelles en analysant l'évolution corrélée de protéines chez plusieurs génomes. La présence ou l'absence simultanée de deux protéines données dans un génome permet de prédire que ces deux protéines sont impliquées dans le même complexe protéique ou la même voie métabolique. Ces protéines sont définies comme liées fonctionnellement.

Il est primordial, dans ce cas, de choisir les génomes de référence de façon adéquate afin de prédire les interactions au plus juste [Sun et al., 2005].

3.1.5.3 Représentation formelle des relations

Des modèles abstraits et structurés représentent les éléments (ou objets) et leurs relations (*e.g.* un graphe décrivant un réseau métabolique). Deux problèmes de représentation se posent alors : le problème de la modélisation des relations ; et celui de la représentation des éléments. Plusieurs projets de recherche se sont développés pour la représentation formelle des relations (par exemple : KEGG et Reactome[14] pour les voies métaboliques). Mais les références entre ces diverses sources de données restent superficielles du fait du choix technique effectué par les équipes de recherche, mais aussi par le problème de désignation d'un objet. En effet, un objet peut être désigné différemment selon la banque de données, rendant les modèles de données, dans ce cas, peu explicites ou incompatibles. L'utilisation d'une ontologie telle que la GO permet d'établir ces relations entre plusieurs occurrences d'un même élément, dans des banques de données respectant la GO.

Ainsi, l'initiative HUPO-PSI[15] ("Human Proteome Organization Proteomics Standards Initiative") [Hermjakob et al., 2004a, Kerrien et al., 2007] définit des standards pour la représentation des données en protéomique afin de faciliter la comparaison, l'échange et la vérification des données au sein de la communauté scientifique. Un de leurs groupes de travail, HUPO-PSI-MI ("HUPO-PSI Molecular Interaction") est dédié aux interactions moléculaires. Il cherche à améliorer, d'une part, l'annotation et la représentation de ces interactions après leur publication et d'autre part, leur accessibilité pour les utilisateurs. Pour cela, HUPO-PSI-MI a développé les informations minimales nécessaires pour décrire une interaction [Orchard et al., 2006] selon le format d'échange de données élaboré par les membres du

[14]Ressource de données pour les voies métaboliques et réactions chez l'Homme : http ://www.reactome.org.
[15]HUPO-PSI : http//www.psidev.info/.

groupe. La diversité des domaines scientifiques dont proviennent les membres du groupe permet une véritable réflexion sur les besoins et les attentes des biologistes et bio-informaticiens. HUPO-PSI fait partie de l'organisation HUPO ("Human Proteomes Organization"), créée en 2001 suite à l'annotation du génome humain [Venter et al., 2001], et dont la mission est de définir et promouvoir la protéomique à travers une coopération et des collaborations internationales en favorisant le développement de nouvelles technologies, techniques et formations afin de mieux comprendre les maladies chez l'homme [consortium, 2005].

La représentation formelle des relations fait l'objet de recherches [Batt, 2001] que nous n'aborderons pas dans ce document.

L'annotation relationnelle requiert les connaissances apportées par l'annotation syntaxique et fonctionnelle. Les méthodes expérimentales utilisées pour l'annotation relationnelle nécessitent un délai de mise en œuvre supplémentaire quand elles ont besoin de synthétiser des sondes (ADN ou ARN) ou de cibler des séquences particulières, en supposant que la technique est maîtrisée et ajustée à l'organisme étudié. Les méthodes prédictives, quant à elles, peuvent donner leurs premiers résultats dès que l'annotation syntaxique et fonctionnelle est disponible. Étant donné leur nature prédictive, elles doivent être ensuite confirmées par les méthodes expérimentales.

Par ailleurs, la complétude de l'annotation relationnelle est difficile à atteindre, contrairement à l'annotation syntaxique. D'une part, l'annotation relationnelle est réalisée, petit à petit, par des équipes de recherche, impliquées ou non dans le projet initial d'annotation génomique de l'organisme étudié. Ces équipes se concentrent en général sur un thème de recherche (voie métabolique, transport, famille de gènes, processus biologique...). Les résultats sont ensuite intégrés dans les banques de données, généralistes et spécialisées. Seule la banque de données dédiée à un seul organisme peut rassembler l'ensemble des informations liées à l'annotation relationnelle, telle que la banque SGD pour *S. cerevisiae*. La collecte des données est alors permanente et importante. Elle requiert aussi une expertise humaine afin d'être de bonne qualité.

D'autre part, les techniques expérimentales actuelles ne permettent pas de mettre en évidence toutes les interactions. Cela est dû aux conditions expérimentales qui peuvent fausser les IPP, aux limites de leur seuil de détection, mais aussi à la nature même de certaines interactions trop fugaces pour être observées.

Grâce au développement de l'informatique et plus précisément de la bio-informatique, l'annotateur dispose d'outils d'aide à l'annotation et de méthodes d'analyses de résultats expérimentaux et prédictifs adaptés aux trois phases de l'annotation. Ces méthodes et outils bio-informatiques se dont développés parallèlement au progrès de l'annotation génomique et du besoin d'analyses rapides et de plus en plus automatisées des données expérimentales produites en grande quantité.

3.2 Stratégies bio-informatiques utilisées pour l'annotation génomique

Après avoir rappelé l'intérêt de la bio-informatique pour l'annotation génomique, nous présentons, à travers les deux premiers grands projets d'annotation de génomes eucaryotes que sont le nématode et la levure, les problèmes auxquels ont dû faire face les annotateurs, ainsi que les méthodes bio-informatiques utilisées. Le travail fourni par ces projets a permis d'améliorer la stratégie et les techniques d'une annotation génomique de plus en plus rapide, homogène et automatisée.

3.2.1 Apport de la bio-informatique pour l'annotation

Le développement de la bio-informatique s'est fait naturellement, dès les années 1950-1960, lorsque les scientifiques ont dû faire face à des problèmes pratiques d'analyse de données biologiques. Par exemple, il n'est pas concevable, d'un point de vue de délai, de rechercher manuellement les séquences codantes sur un génome, ou de comparer des séquences entre elles.

Utilisant les potentialités de l'informatique (automatisation de tâches, modèles théoriques, algorithmes, programmes, bases de données...), la bio-informatique a contribué, entre autres, à l'invention d'algorithmes de recherche et d'analyse de séquences.

Ainsi, les experts biologiques bénéficient d'une aide à l'annotation grâce à diverses méthodes d'analyse du génome mais également d'environnements informatiques permettant de meilleures organisation et homogénéisation de leur travail.

3.2.1.1 Aide à l'annotation

Nous pouvons distinguer deux niveaux d'aide à l'annotation :
- les méthodes d'analyse du génome telles que la prédiction de gènes ou bien la recherche de motifs particuliers,
- et les plate-formes d'annotation, proposant un environnement informatique aux annotateurs, et intégrant les résultats des méthodes d'analyses précédentes.

Consécutivement au séquençage des génomes bactériens, premiers à être séquencés, des logiciels d'aide à l'annotation ont été développés. Ils ont ensuite été adaptés pour être utilisés chez les eucaryotes dont les génomes sont plus complexes (présence d'introns) et nécessitent des méthodes prédictives de gènes adéquates. Le nombre de méthodes et de plate-formes développées étant conséquent, nous ne citerons que celles qui nous semblent importantes de par leur antériorité, leur succès d'utilisation ou le contexte dans lequel elles ont été développées.

Les méthodes d'analyse Les méthodes pour la prédiction de gènes les plus couramment utilisées se basent sur des méthodes *ab initio* telles que la détection du potentiel codant d'une séquence, réalisé par le programme GeneMark [Borodovsky and McIninch, 1993]. Basés sur les chaînes de Markov, GeneMark et Glimmer [Salzberg et al., 1998], une autre méthode de prédiction de gène, ont été initialement créés dans un contexte procaryote avant d'être

utilisés chez les eucaryotes. Les méthodes courantes de prédiction d'introns chez les eucaryotes se basent elles-aussi sur les chaînes de Markov telles AUGUSTUS [Stanke, 2003] ou JigSaw [Allen and Salzberg, 2005].

Les méthodes par recherche de séquences homologues (logiciel BLAST [Altschul et al., 1990]) ou de motifs (ScanProsite [Gattiker et al., 2002]) aident également l'annotateur dans son identification de séquences codantes.

Le logiciel Exogean ("EXpert On GEne Annotation", développé à l'École Normale supérieure de Paris) [Djebali et al., 2006] permet la prédiction de gènes eucaryotes par annotation automatique. Sa participation au projet EGASP (cf. § 3.2.1.2) l'a classé comme meilleure méthode d'annotation automatique parmi les vingt en compétition.

Les outils bio-informatiques facilitent ainsi l'annotation génomique à plusieurs points de vue dont les plus importants sont : (i) la recherche d´ORF, de CDS et d'intron ; (ii) la recherche de motifs transcriptionnels et traductionnels ; (iii) la recherche de séquences homologues dans les banques de données ; (iv) les recherches de domaines protéiques ; (v) et la phylogénie moléculaire (études de l'évolution des gènes, ARN et protéines).

L'annotation d'un génome faisant appel à diverses méthodes de prédiction (prédiction de séquences codantes, promoteurs, pseudogènes...), plusieurs méthodes sont intégrées au sein d'une plate-forme d'annotation. Les annotateurs disposent ainsi aisément, grâce à une interface graphique accessible par internet, des différents résultats de ces méthodes.

Les plate-formes d'annotation L'un des premiers logiciels d'annotation automatique est MAGPIE ("Multipurpose Automated Genome Project Investigation Environment") [Gaasterland and Sensen, 1996b, Gaasterland and Sensen, 1996a], utilisé pour les organismes procaryotes. Les annotateurs pouvaient intervenir sur le système pour préciser leurs préférences et l'importance des analyses les unes par rapport aux autres. Ils pouvaient également modifier les annotations obtenues automatiquement en cas de désaccord, ce qui donnait une annotation automatique mais néanmoins vérifiée par un expert biologiste. Puis en 1999, l'environnement informatique Imagene [Médigue et al., 1999] rassemblaient, dans un même modèle orienté objet, les données biologiques issues d'un projet de séquençage de génome bactérien et les méthodes d'analyses des séquences en vue de l'annotation manuelle de ce génome. Plus récemment, plusieurs projets d'annotation manuelle de génomes procaryotes utilisent les systèmes MaGe ("Magnifying Genome", développé au Génoscope) [Vallenet et al., 2006] et AGMIAL ("Analyse de Génomes Microbiens d'Intérêt Agro-aLimentaire", développé à l'INRA de Jouy-en-Josas) [Bryson et al., 2006].

Quant à l'annotation de génomes eucaryotes, elle bénéficie de plate-formes identiques à celles pour les procaryotes, telle Artemis [Rutherford et al., 2000], outil d'annotation et de visualisation de génome développé par l'Institut Sanger pour l'annotation de petits génomes, procaryotes ou eucaryotes. MAGPIE, à l'origine pour les procaryotes, s'est enrichi entre autres de la méthode GENSCAN [Burge and Karlin, 1997] pour la détection d'introns afin d'être applicable chez les eucaryotes. Le système a alors était rebaptisé

EGRET [Gaasterland et al., 2000]. En 2002, l'institut Sanger et le consortium du centre du génome de la drosophile (université de Berkeley) ont développé Apollo [Lewis et al., 2002], autre outil d'aide à l'annotation spécifique des génomes eucaryotes de grande taille.

L'ensemble de ces systèmes d'aide à l'annotation permet d'intégrer des analyses bio-informatiques couramment utilisées pour l'annotation tels que les résultats d'alignement FASTA ou BLAST. De plus, ils gèrent divers formats de données (*e.g.* FASTA, EMBL) repris par les banques de données, facilitant ainsi l'échange des données entre ces banques et le résultat de l'annotation. Certains systèmes permettent en plus une annotation relationnelle, tel le système d'annotation MaGe qui s'intéresse à la reconstruction et la visualisation des voies métaboliques de l'organisme.

3.2.1.2 Évaluation des méthodes prédictives pour l'annotation : le projet EGASP

La diversité et la multiplicité des méthodes d'aide à l'annotation ne cesse de s'accroître au fil du temps. Un laboratoire de recherche qui voudrait utiliser une méthode plutôt qu'une autre, intégrée dans un système d'aide à l'annotation, devra ainsi étudier l'existant en considérant ses besoins et les contraintes liées à son projet d'annotation. L'annotation de génome étant devenue une étape obligatoire dans la découverte des connaissances biologiques sur les organismes, l'état de l'art sur les moyens informatiques d'identification des gènes est en fait régulièrement publié dans les revues de ce domaine telles que Trends in Genetics [Fickett, 1996], Bioinformatics [Baldi et al., 2000], Nucleic Acids Research [Mathé et al., 2002], Genomics Proteomics Bioinformatics [Wang et al., 2004], Genome Research [Brent, 2005] et Journal of Microbiology [Do and Choi, 2006].

Les auteurs de méthodes prédictives ou de plate-formes d'annotation comparent parfois leur système à ceux existants, tel Bryson *et al* pour leur système AGMIAL. Cela permet ainsi d'avoir un premier aperçu des diverses possibilités offertes pour un nouveau projet d'annotation génomique. Mais cette présentation adopte le point de vue des auteurs du système, qui souhaitent montrer que leur système apporte une nouvelle contribution au domaine.

Aussi, la meilleure façon d'évaluer les différents systèmes demeure leur utilisation sur un jeu de données standard afin de comparer les résultats. Mais cela ne peut être réalisé que sur des résultats obtenus par annotation automatique, l'expertise humaine n'étant pas reproductible ni d'un individu à un autre ni pour un seul homme (l'expert acquiert de l'expérience au fil de son travail). Par la suite, les annotateurs peuvent, s'ils le souhaitent, raffiner cette première annotation.

Comme nous l'avons vu en introduction, deux projets ont permis ainsi d'évaluer des logiciels de prédictions pour l'annotation automatique sur un jeu de données test. Le projet GASP ("Gene Annotation aSsessment Project") [Reese et al., 2000] comparait les résultats de méthodes prédictives d'éléments géniques sur un jeu de données de drosophile, dont les données contrôles étaient annotées manuellement. Ce même principe fut repris en 2006 par le projet EGASP ("ENCODE GASP") [Guigo et al., 2006] sur un jeu de 44 régions humaines. Ces régions, soit ∼1% du génome humain, ont été défi-

nies comme représentatives du génome humain par le projet ENCODE ("ENCyclopedia Of DNA Elements") [ENCODE Project consortium, 2004]. Puis le consortium GENCODE [Harrow et al., 2006] a annoté manuellement ces séquences de façon à obtenir une annotation de très haute qualité. Les équipes engagées entraînaient les logiciels, répartis en diverses catégories (méthodes utilisant tout type de données, méthodes *ab initio* pour un génome, prédictions seulement des exons ...), sur 13% des régions ENCODE. Les logiciels évalués permettent la détection des éléments principaux, *i.e.* les gènes codant les protéines, avec Exogean [Djebali et al., 2006] et les logiciels de détection d'intron [Allen et al., 2006], les pseudogènes [Zheng and Gerstein, 2006], les promoteurs [Bajic et al., 2006]. Certains, tels AceView [Thierry-Mieg and Thierry-Mieg, 2006] et AUGUSTUS [Stanke, 2003], basent leur prédiction sur des résultats expérimentaux.

3.2.1.3 Contraintes liées au contexte d'un projet d'annotation

Il apparaît que la stratégie adoptée pour l'annotation génomique est fonction des contraintes liées à son contexte qui comprend : l'architecture du génome séquencé ; les données biologiques disponibles ; l'état d'assemblage des séquences génomiques ; et le nombre de personnes impliquées dans le projet d'annotation.

Architecture du génome séquencé L'organisation d'un génome procaryote est plus simple que celle d'un eucaryote. Tout d'abord, le génome procaryote est plus petit et plus compact, *i.e.* la molécule d'ADN génomique a une densité en gènes plus élevée que chez les eucaryotes (95% du génome du bacille *E. coli* est transcrit contre seulement 1,8% pour l'homme). De plus, les gènes procaryotiques n'ont pas d'intron, ce qui facilite la détection des ORF. Chez les procaryotes et certains eucaryotes, certains gènes sont regroupés en opéron[16] : ils sont très proches les uns des autres.

Données biologiques disponibles Les annotateurs peuvent disposer de séquences obtenues lors d'études antérieures au séquençage, les aidant dans leur travail d'annotation (syntaxique, fonctionnelle et relationnelle). Par exemple, l'annotation d'un génome apparenté au génome en cours d'annotation facilite la prédiction de séquences homologues. Les études à grande échelle du séquençage du transcriptome par l'obtention de marqueurs de séquences exprimées[17], ou EST ("Expressed Sequence Tag"), permettent également de connaître une partie des gènes exprimés.

Assemblage des séquences génomiques La qualité du génome à annoter dépend de l'état de finition de l'assemblage des séquences génomiques. Dans le meilleur des cas, le génome est disponible sous sa forme finale : les chromosomes sont entièrement définis. Dans le cas

[16] Un opéron est une unité d'expression génétique qui compte un ou plusieurs gènes et des séquences régulatrices (promoteur et opérateur) qui régulent leur transcription.

[17] Les ARNm d'une cellule sont rétro-transcrits en ADN complémentaires, ou ADNc. Ces ADNc sont séquencés à leurs extrémités : ces séquences partielles d'ADNc sont appelées les EST. L'information est partielle mais suffisante pour caractériser chaque ADNc et, par conséquent, le gène dont est issu l'ARNm.

où des séquences partielles de chromosomes, les contigs et super-contigs, sont disponibles, il est néanmoins possible d'en faire l'annotation du fait de leur taille suffisante (plusieurs centaines de milliers de bases). Les séquences qui feraient le lien entre ces contigs ou super-contigs et les chromosomes, ont en fait une structure particulière (notamment la présence de séquences répétées), rendant difficile leur séquençage. Or ces trous dans la séquence génomique représentent autant de régions codantes potentielles qui ne pourront être prises en compte dans l'annotation.

Participants au projet d'annotation L'annotation d'un génome est un travail long et fastidieux. Celui-ci est réparti entre les différents membres du projet compétents pour l'annotation. Plus un projet compte d'annotateurs, plus vite le travail d'annotation est réalisé. La communication entre annotateurs, qu'ils soient distants géographiquement ou non, est facilité par la mise en place d'une interface de gestion d'annotation, accessible par internet. L'utilisation d'une plate-forme commune permet également aux annotateurs d'homogénéiser leur travail (sources de données et paramètres communs, procédure standardisée d'annotation...).

3.2.2 Premiers grands projets d'annotation génomique

Les stratégies d'analyse des séquences ont été mises en place lors des premiers projets d'annotation génomique apparus en 1995. Ces projets portaient sur l'annotation de génomes bactériens. Ils ont contribué à la mise au point et la validation des procédures d'annotation manuelle, semi-automatique et automatique.

En 2002, une centaine de micro-organismes procaryotes étaient entièrement séquencés et annotés, parmi lesquels les bactéries *Mycoplasma genitalium* [Fraser et al., 1995], *Haemophilus influenza* [Fleischmann et al., 1995], *E. coli* [Blattner et al., 1997].

En 2002, les premiers projets d'annotation de génomes eucaryotes concernent huit espèces, presque toutes des organismes modèles : la levure *S. cerevisiae* [Goffeau et al., 1996], le nématode *C. elegans* [Berks and the *C. elegans* GMSC, 1995, *C. elegans* Sequencing Consortium, 1998], la drosophile *D. melanogaster* [Adams et al., 2000], les plantes *A. thaliana* [Initiative, 2000] et *Oryza sativa* (le riz[18]) [Goff et al., 2002, Yu et al., 2002], la souris *M. musculus* [Waterston et al., 2002], et l'homme *Homo sapiens sapiens* [Venter et al., 2001].

Nous avons choisi de présenter les stratégies bio-informatiques des deux premiers et principaux projets d'annotation de génomes eucaryotes, précédant le projet Génolevures, en considérant les problématiques de rapidité et qualité énoncées en introduction. Ces problématiques sont communes à tout projet d'annotation.

L'annotation, manuelle dans tous les cas, ne comporte que les aspects syntaxiques et fonctionnels. L'annotation relationnelle a pris une part importante dans l'annotation génomique à partir des années 2000, lorsque les techniques d'étude des interactions entre éléments cellulaires ont pu être appliquées à grande échelle, permettant alors de concevoir un génome comme un système biologique entier.

[18]Deux souches de riz ont été séquencées.

3.2.2.1 Le nématode

Dans les années 1960, Sydney Brenner [19] choisit le nématode *C. elegans* comme organisme modèle pour l'étude du développement des cellules animales, en particulier les cellules du système nerveux, et la physiologie de l'organisme. Depuis, le nématode a permis de grandes avancées sur la formation du système nerveux [Strange, 2006].

L'annotation de son génome (6 chromosomes constituant 100 Mpb, 20 049 gènes) a commencé en 1985 par la collaboration de deux laboratoires (Saint-Louis aux États-Unis, et Cambridge au Royaume-Uni) et s'est achevée en 1998 [*C. elegans* Sequencing Consortium, 1998]. Le délai d'obtention de la séquence génomique imposa le rythme d'annotation du projet.

En 1995, 20% du génome était annoté [Berks and the *C. elegans* GMSC, 1995]. Le programme GeneFinder [Green and Hillier, ults] a servi à l'identification des ORF et des sites potentiels donneurs et accepteurs d'épissage. Les recherches de similarités de séquences des motifs protéiques, gènes d'ARNt et EST disponibles avec les données de banques publiques, ont été réalisées en utilisant BLAST (BLASTx pour les similarités entre protéines et BLASTn pour celles entre nucléotides) [Altschul et al., 1990].

De plus, l'ensemble des séquences disponibles sont comparées entre elles afin d'identifier des familles de gènes. L'ensemble des données est alors stocké dans la base de données ACeDB [Durbin and Thierry-Mieg, 1991].

Alors que le séquençage était essentiellement fini en 1996, l'annotation de *C. elegans* continua pendant encore deux ans, au fur et à mesure que de nouvelles informations biologiques paraissaient et que les outils d'annotation de séquences devenaient plus performants. Chaque séquence fut soumise à une suite d'analyses automatiques afin d'identifier les protéines potentielles avec GeneFinder, les gènes d'ARNt avec tRNAscan [Fichant and Burks, 1991] et tRNAscan-SE [Lowe and Eddy, 1997], les similarités aux EST disponibles et les autres protéines avec le programme WU-BLAST[20] [Gish, 1995], les familles de séquences répétées et répétitions locales.

Chez le nématode, outre le problème d'épissage et de faible densité des gènes, les processus d'épissage en *trans*[21] et d'organisation en opéron ont rendu la détection des séquences codantes plus complexe. Cette difficulté a été résolue, dans la majorité des cas, par l'analyse d'EST et d'ADNc obtenus expérimentalement, ainsi que la comparaison avec des séquences génomiques d'un nématode apparenté.

Lors de la parution, en 1998, de l'annotation de *C. elegans*, le consortium notait que ce projet avait contribué au développement des techniques expérimentales et informatiques pour le séquençage et l'analyse des données. Il est vrai que GeneFinder, tRNAscan et tRNAscan-SE ont été développés et/ou améliorés lors de ce projet. De ce fait, les annotations ont été obte-

[19] S. Brenner reçut le prix Nobel de Médecine en 2002, en compagnie de Robert Horvitz et John Sulston, pour leurs travaux complémentaires sur la mort programmée de la cellule, menés sur le nématode.

[20] WU-BLAST permet un alignement local entre deux séquences en autorisant des trous selon des paramètres statistiques.

[21] L'épissage en *trans* met en jeu deux ARN pré-messagers ayant au moins un intron. Ces ARNpré-m s'échangent une partie de leur séquence au niveau d'une séquence intronique, obtenant par la suite deux ARNm. L'épissage d'intron(s) portant sur un seul ARN pré-messager (forme d'épissage classique), est appelé épissage en *cis*.

nues selon des procédures et des paramètres hétérogènes et spécifiques à chaque laboratoire intervenant dans le projet.

3.2.2.2 La levure

Comme nous l'avons vu § 2.1.3.1, la levure *S. cerevisiae* est l'organisme de réference pour l'étude des mécanismes cellulaires de la cellule eucaryote. Le séquençage et l'annotation finale de son génome (16 chromosomes constituant 12,3 Mbp, 6 609 ORF[22]) ont été publiés en 1996 [Goffeau et al., 1996] mais le projet a démarré en 1992, lors du séquençage du chromosome III (315 kb pour 182 ORF) [Oliver et al., 1992] de la levure, premier chromosome entier paru, tous organismes eucaryotes confondus.

L'annotation manuelle du génome de *S. cerevisiae* est le fruit d'une collaboration de plus de 600 personnes, répartis dans de nombreux laboratoires à travers le monde (Allemagne, Belgique, Canada, États-Unis, France, Japon, Royaume-Uni, Suisse). Chacun des laboratoires participant avait en charge le séquençage et l'annotation d'un chromosome ou d'une région chromosomique. Les résultats d'analyse étaient ensuite publiés au fur et à mesure. Par conséquent, chaque projet a mis en place sa propre stratégie d'annotation. Mais une étroite et forte collaboration entre partenaires a permis la publication de résultats basés sur les mêmes standards et normes, grâce à l'utilisation d'internet et des techniques informatiques de l'époque.

Les techniques de séquençage automatisé n'ayant pas encore cours, les principaux obstacles du projet d'annotation de *S. cerevisiae* furent le délai et le coût nécessaires à l'obtention du séquençage entier de ce génome, soit 4 ans, malgrè la mobilisation et la motivation des nombreux laboratoires impliqués.

Exemple de l'annotation du chromosome VIII En 1994, l'équipe de Johnston [Johnston et al., 1994], responsable du chromosome VIII (563 kpb), fit appel à des méthodes bio-informatiques qui avaient déjà fait leurs preuves lors de l'annotation de *C. elegans*. L'annotation commença dès que des séquences d'ADN génomiques de 30 à 40 kpb furent disponibles. Ces séquences ont été comparées avec les banques de séquences nucléiques et protéiques disponibles (SwissProt, GENEPEP, PIR), avec les programmes BLASTn (pour la recherche de similarités nucléotidiques) et BLASTx (pour la recherche de similarités protéiques) de l'outil BLAST [Altschul et al., 1990]. Les ARNt ont été recherchées avec le logiciel tRNAScan [Fichant and Burks, 1991]. Elles ont ensuite été assemblées et annotées avec AScDB, une version du programme ACeDB [Durbin and Thierry-Mieg, 1991] utilisée pour l'annotation manuelle de *C. elegans* et modifiée pour prendre en compte les données de *S. cerevisiae* disponibles. L'annotation syntaxique porta sur l'identification de tous les ORF d'au moins 100 codons, codons start et stop inclus. De nos jours, ce seuil minimal fixé apparaît bien supérieur à la taille minimale des protéines détectées chez *S. cerevisiae* mais aussi chez d'autres organismes. Si cela n'a pas été fait comme par exemple avec *S. cerevisiae* [Blandin et al., 2000], il faudrait refaire l'annotation des génomes avec des paramètres plus proches de la réalité biologique afin d'en améliorer la qualité. Le programme GeneFinder [Green and Hillier, ults],

[22]Données de SGD, 10/09/2007.

adapté aux données de *S. cerevisiae*, a aidé aux choix des gènes prédits. En cas de chevauchement de gènes, le gène le plus long et ayant une homologie avec un autre gène, était choisi. Le premier codon start trouvé dans une ORF était systématiquement sélectionné comme début du gène. Les sites d'épissage étaient également recherchés quand un intron était suspecté d'être présent et permettait d'obtenir un gène entier. Le point de branchement, défini par sa séquence `TACTAAC`, était recherché dans une zone de 5 à 134 bases en amont du site accepteur. Ces caratéristiques très spécifiques (séquence et distance) s'appuyaient sur des observations expérimentales antérieures mais ne laissaient aucune flexibilité dans la recherche d'intron. Celle-ci était très spécifique mais peu sensible. Les analyses ultérieures du génome de *S. cerevisiae* ont révélé que ces choix restreignaient la détection des introns : le motif du point de branchement, choisi à cette époque, est présent dans 85% des introns prédits à ce jour. Ces critères de recherche d'intron mettent ainsi en évidence l'absence de complétude de l'annotation parue. Des éléments mobiles furent également recherchés en comparant les séquences génomiques avec une séquence représentative de chaque élément à l'aide des programmes BLASTn et FASTA.

La comparaison des ORF avec les protéines présentes dans les banques de séquences a également mis en évidence des erreurs potentielles de séquençage (la présence d'un décalage du cadre de lecture, ou "frameshift", dans des zones conservées entre des séquences d'espèces différentes).

L'annotation de *S. cerevisiae* de 1996 a été corrigée et améliorée jusqu'en 2000 ponctuellement, grâce à des études sur le transcriptome [Velculescu et al., 1997], le reséquençage de certaines régions ou des études spécialisées sur les introns, les gènes d'ARNt, les transposons. En 2000, le projet Génolevures 1 a permis la réannotation manuelle [Blandin et al., 2000] du génome de *S. cerevisiae* par comparaison avec les 13 nouvelles espèces partiellement séquencées et annotées. Génolevures 1 a ainsi mis en évidence 50 nouveaux gènes de *S. cerevisiae*. Quarante-huit de ces gènes sont prédits codant une protéine d'une taille inférieure à 100 acides aminés ; deux gènes, prédits codant une protéine de plus de 100 aa, ont chacun un intron. Par ailleurs, la comparaison entre ces différentes levures a mis en évidence d'autres motifs d'épissage intronique chez *S. cerevisiae* que ceux connus jusqu'alors. Le tableau 3.1 indique les différences d'annotation entre la parution initale de l'annotation en 1996 et celle à ce jour. Le taux d'erreur concernant l'identification des ORF est de 5% ce qui reste acceptable. Le taux d'erreur de détection d'intron dans un gène est de 13% : la première annotation avait manqué alors de nombreux faux négatifs.

L'annotation de *S. cerevisiae* aura ainsi duré huit ans avant d'atteindre une bonne qualité dans l'identification de l'ensemble des éléments, codants ou non, du génome, chaque équipe participant au projet ayant réalisé son travail d'annotation selon ses propres critères. Mais cette annotation n'a donné qu'un catalogue de gènes et éléments non codants, auquel les expériences biologiques étudiant les interactions entre éléments cellulaires ont été intégrées, donnant ainsi un début d'annotation relationnelle. Cette annotation relationnelle s'est réellement enrichie lors des premières expériences à grande échelle d'interaction protéine-protéine dès 2001 [Ito et al., 2001, Ho et al., 2002, Gavin et al., 2002].

Nous pouvons également noter que la partie la plus longue du projet est le séquençage

TAB. 3.1 – Comparaison de l'annotation de *S. cerevisiae* entre 1996 et 2007.

	1996 [Goffeau et al., 1996]	2008 (SGD, 01/24/08)
ORF	6 275	6 576
ORF hypothétiques	5 885	5 777
ORF questionnables	390	804
ARNt	275	275
ARNsn	40	6
ARNr	140	25
total	6 730	6 882
intron dans gène	220	253

Les données concernent uniquement le génome nucléaire. Les ORF questionnables n'ont pas été observés sous forme protéique, contrairement aux ORF hypothétiques dont des produits de traduction ont été mis en évidence. Les ORF hypothétiques et questionnables constituent l'ensemble des ORF prédits. Les ORF hypothétiques pour 2008 regroupent 4 650 ORF vérifiés et 1 122 ORF non caractérisés. À ce jour, il y a 274 ARNt fonctionnels et un ARNt non fonctionnel. Des changements de nomenclature dans l'annotation expliqueraient la grande diminution du nombre d'ARNsn et ARNr.

du génome lui-même, dû aux techniques encore peu performantes et relativement chères à l'époque. Ainsi, près de 55% du génome a été séquencé par des laboratoires européens de recherche académique. De ce fait, ces laboratoires n'étaient pas sous la pression de centres de séquençage pour la publication des séquences. De plus, face à la lenteur d'obtention des séquences, les annotateurs ne se sont pas retrouvés face à une quantité importante de séquences à annoter d'un coup, comme c'est le cas actuellement.

Comme pour le projet d'annotation de *C. elegans*, les logiciels GeneFinder, tRNAscan et tRNAsan-SE ont été améliorés lors de ce projet. De ce fait, les annotations ont été obtenues selon des procédures et des paramètres hétérogènes et spécifiques à chaque laboratoire intervenant dans le projet.

3.2.3 Projets de génomique comparée

Par définition, un projet de génomique comparée porte sur l'analyse de l'intégralité d'au moins deux génomes. Jusqu'en 2002, de par la nature des données disponibles, ces études de génomique comparative se retrouvent face à deux limites. Soit elles sont restreintes à la comparaison de génomes complets mais d'espèces distantes phylogéniquement. Par exemple, en 1997 la distribution des ORF est comparée chez cinq organismes (trois bactéries, le nématode et la levure) [Coissac et al., 1997]. En 2000, les génomes complets du nématode, de la levure et de la drosophile, sont comparés selon les points de vue des processus de la cellule, du développement et de l'évolution [Rubin et al., 2000]. Soit ces études de génomique comparative sont restreintes à la comparaison d'espèces apparentées mais portant sur quelques traits de caractères ou n'ayant que peu de données. Par exemple, la conservation de quatre gènes présentant un domaine structural particulier chez l'homme et les grands singes, suggère un

événement de duplication ou de retrotranscription pour ces gènes-là avant la séparation des branches phylogéniques de ces deux genres [Villa et al., 1996]. Certaines études arrivent cependant à trouver une approche médiane. Par exemple, la comparaison entre les génomes de la souris et de l'homme se basent sur 1196 gènes orthologues [Makalowski et al., 1996].

Cependant, toutes ces études n'ont pas une portée aussi limitée qu'il semble paraître car elles émettent les premières hypothèses concernant l'évolution des génomes et leur spéciation. à partir de 2002, la mise à disposition de génomes nouvellement annotés permet aux études de génomiques comparatives d'être plus complètes et précises dans les hypothèses liées aux mécanismes de l'évolution et du fonctionnement des organismes [Karlin et al., 2003].

3.2.4 Cohérence de l'annotation

Comme nous l'avons vu dans le chapitre introductif (cf. p. 6), les informations biologiques extraites de l'annotation d'un génome doivent permettre de mieux comprendre le fonctionnement de l'organisme et d'extrapoler celui d'autres organismes apparentés. Étant donné sa réutilisation à des fins biologiques et bioinformatiques, l'annotation doit être la plus juste possible. Nous présentons ici deux projets qui intègrent des outils d'amélioration, à leur niveau, la qualité de l'annotation d'un génome.

3.2.4.1 Détection d'annotation protéiques erronées : le système Xanthippe

Le système Xanthippe [Wieser et al., 2004], développé à l'EBI ("European bioinformatics Institute") détecte les annotations erronées. Comme nous l'avons vu en introduction, l'annotation automatique peut propager les annotations de séquence homologue en séquence homologue. Par exemple, une règle d'annotation automatique, telle que le proposent RuleBase [Biswas et al., 2002] et Spearmint [Kretschmann et al., 2001], peut imposer qu'à toute protéine ayant un domaine de liaison à l'ADN, soit attribué le mot-clef "protéine nucléaire". Cela peut être vrai si la protéine appartient à un organisme eucaryote, pourvu d'un noyau, mais pas pour une protéine d'origine procaryote. Cette règle de propagation du mot-clef est née de l'apprentissage du système avec un jeu de données. Si le jeu de données de départ est biaisé (ici, le jeu de données n'est composé que de protéines eucaryotes), les résultats seront faussés. Dans notre exemple, l'absence de cas vrai négatif ("une protéine procaryote ayant un domaine de liaison à l'ADN n'est pas une protéine nucléaire") ne signale pas au système d'annotation le conflit existant.

Le système Xanthippe se base sur un mécanisme d'exclusion simple et une approche par arbre de décision basé sur l'algorithme de data mining C4.5 [Quilan, 1993]. Ce système de détection d'erreur s'applique sur les annotations protéiques issues de l'annotation automatique, non vérifiées manuellement par un expert.

3.2.4.2 Utilisation de règles négatives

Une approche différente et singulière de la détection d'erreurs dans l'annotation est celle utilisant des règles négatives telle que le propose Artamonova *et al.* [Artamonova et al., 2007]. Nous reprenons le contexte de l'exemple cité § 3.2.4.1 ci-dessus avec la protéine ayant un

domaine de liaison à l'ADN. Dans ce contexte, une règle positive est "la protéine a un domaine de liaison à l'ADN alors elle est nucléaire" : la relation existante entre la condition et la conséquence est positive. À l'opposé, une règle négative est "la protéine a un domaine de liaison à l'ADN alors elle n'est pas d'origine bactérienne".

Les auteurs ont ici fait appel à l'algorithme Apriori [23] [Agrawal et al., 1993] pour générer les règles négatives à partir des annotations automatiques de la banque de données PEDANT [Frishman et al., 2001, Riley et al., 2007]. L'application de ces règles aux protéines annotées permet de détecter les combinaisons d'annotations incompatibles.

Les approches utilisées par ces deux systèmes ne satisfont pas les critères que nous nous sommes fixés (complétude et absence d'erreur, cf. § 1) et ce, pour deux raisons. La première concerne la considération individuelle des protéines par la vérification de la cohérence de l'annotation. La seconde raison concerne la vérification de seulement l'annotation fonctionnelle des protéine. Ceci est une conséquence indirecte de la première. En effet, si les protéines sont considérées individuellement, il ne peut y avoir de vérification sur l'absence de chevauchement entre les séquences, ou la présence des fonctions essentielles pour un organisme, étant donné que cela nécessite une vue d'ensemble de l'annotation du génome.

Le projet Génolevures nous permet de disposer d'un cadre adéquat sur les deux points de vue. D'une part, ce projet réalise l'annotation manuelle de plusieurs génomes, ce qui nous donne les jeux de données pour l'application des règles de cohérence. D'autre part, ce projet se consacre à l'étude des levures, dont fait partie l'organisme modèle *S. cerevisiae* et dont l'utilisation en biologie cellulaire est relativement facile.

3.3 Présentation du projet Génolevures

3.3.1 Le projet Génolevures

Créé en 1998, le projet Génolevures est une étude à grande échelle de génomique comparée concernant les levures de la classe Hémiascomycète. Le consortium Génolevures bénéficie de la structure Groupe de Recherche (GDR) du Centre National de Recherche Scientifique (CNRS) et est coordonné par le Professeur Jean-Luc Souciet, de l'Université Louis Pasteur de Strasbourg. Il se compose de laboratoires de l'Institut Pasteur (Paris), de l'Institut National Agronomique de Paris-Grignon (INA-PG), des Universités Bordeaux 1 et 2, Claude Bernard (Lyon), Paris-Sud (Orsay), Pierre et Marie Curie (Paris 6), Louis Pasteur (Strasbourg), l'Institut Curie (Paris), le Centre d'Essai Atomique (Saclay), le Génoscope (Evry), le Génopole Pasteur-Ile-de-France (Paris) et l'Institut des Sciences de la Vie (Louvain-la -Neuve, Belgique).

Le projet Génolevures est le premier projet de génomique comparée à s'intéresser aux mécanismes évolutifs entre les espèces au sein d'une branche homogène d'espèces. L'objectif de Génolevures est de comprendre les mécanismes concernant l'évolution moléculaire des génomes de levures. Ceci passe par la définition des gènes spécifiques de la branche phylogénique

[23] Le principe de l'algorithme Apriori consiste d'abord à rechercher les ensembles d'éléments fréquents puis de construire des règles à partir de ces ensembles, qui satisfont les contraintes fixées par l'utilisateur.

et ceux spécifiques d'une espèce, la connaissance de la distribution des gènes en familles fonctionnelles, du taux de divergence entre espèces et des mécanismes de réarrangements des chromosomes. Avec leur taille de génome relativement petite (génome en moyenne de 12 Mpb pour 6 000 gènes), ces levures offrent une opportunité unique pour l'étude de l'évolution de génomes eucaryotes, d'après des analyses comparées de plusieurs espèces. En effet, les levures choisies sont à une distance évolutive au moins aussi vaste que celle qui sépare les urochordés marins de l'homme. Cependant, ces organismes sont relativement proches d'un point de vue phylogénique, en terme de physiologie et de morphologie, ce qui constitue un avantage pour leur étude.

Le projet comporte trois phases successives, intégrant de nouveaux aspects de la génomique et des connaissances de pointe. Les trois phases sont déterminées par les données de séquences sur lesquelles les analyses sont réalisées.

Génolevures 1 (1998-2000) [Souciet et al., 2000] disposait de séquences partielles de treize nouvelles espèces de levures comparées à *S. cerevisiae*.

Génolevures 2 (2000-2003) [Dujon et al., 2004] consiste en une analyse plus fine des objectifs de Génolevures 1, car le consortium disposait alors de quatre génomes complets annotés.

Actuellement en cours, Génolevures 3 (2003-2007. . .) se concentre sur la clade des Kluyveromyces, une sous-branche des Hémiascomycètes. L'objectif est l'étude de l'évolution de cinq espèces de Kluyveromyces, réalisable notamment par l'annotation de trois nouveaux génomes.

3.3.2 Levures du projet

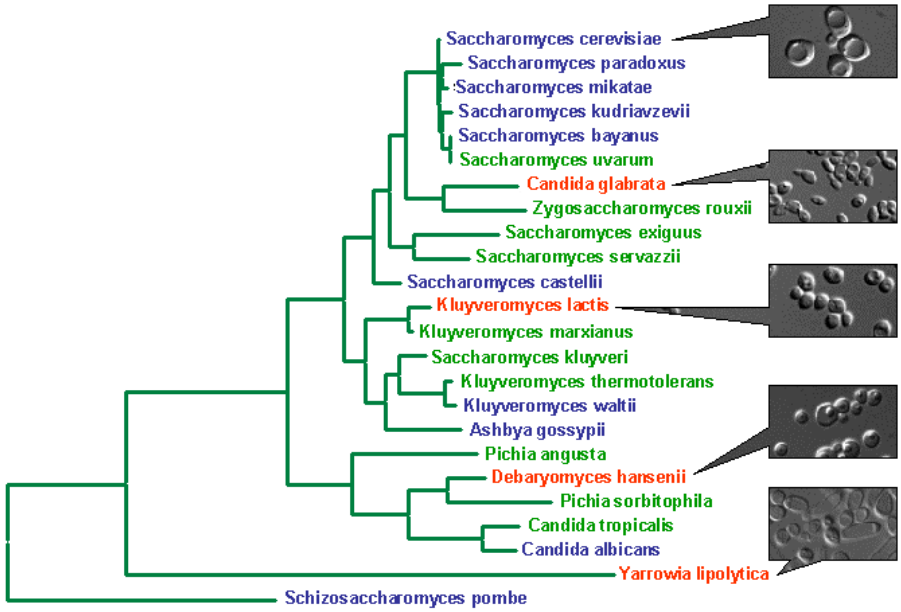
Les treize espèces étudiées lors de la phase Génolevures 1 ont été choisies en fonction, d'une part, de leur représentativité de l'arbre phylogénique des levures hémiascomycètes, et d'autre part, de leurs intérêts en recherche fondamentale, médicale et appliquée.

Le tableau 3.2 présente les caractéristiques des treize levures étudiées dans le cadre de Génolevures 1, ainsi que leurs intérêts les plus notables dans les domaines agro-alimentaire, médical et biotechnologique. Ces espèces ont fait l'objet d'un séquençage aléatoire partiel de leur génome. Selon les espèces, 2 500 ou 5 000 séquences d'ADN génomique de 1 kb ont été annotées manuellement par le consortium, puis étudiées et comparées à *S. cerevisiae*.

Génolevures 2 est le premier grand projet de génomique comparée spécifique d'une branche phylogénique. Suite au séquençage complet des génomes par le Génoscope, le consortium Génolevures a annoté manuellement trois génomes de Génolevures 1 *K. lactis*, *D. hansenii* et *Y. lipolytica*, ainsi que le génome d'une nouvelle levure *Candida glabrata*, pathogène humain plus nocif que *C. tropicalis* (cf. fig. 3.6). Les raisons du choix de ces espèces sont les mêmes que celles citées précédemment. Pour information, les distances évolutives qui séparent *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica* sont respectivement comparables à celles qui séparent l'homme, la souris, le tétraodon (entre *K. lactis* et *D. hansenii*) et la cione (un urochordé marin).

Actuellement en cours, la phase Génolevures 3 se concentre sur la clade homogène des Kluyveromyces en étudiant les levures *Z. rouxii*, *S. kluyveri*, *K. thermotolerans*, *K. lactis*, *Kluyveromyces waltii* et *Ashbya (Eremothecium) gossypii*, un agent pathogène pour le coton. Le consortium Génolevures a séquencé entièrement et annoté de façon semi-

FIG. 3.6 – Phylogénie des levures hémiascomycètes.
Arbre phylogénique calculé par la méthode du maximum de parcimonie avec les gènes de l'ARNr 25S. *Schizossacharomyces pombe* : levure ascomycète, les autres : hémiascomycètes. En vert et rouge : levures partiellement séquencées et annotées du projet Génolevures 1, en rouge : levures de Génolevures 1 entièrement séquencées et annotées lors de Génolevures 2.



automatique *Z. rouxii*, *S. kluyveri* et *K. thermotolerans*. *K. waltii* et *A. gossypii* ont été séquencés et annotés respectivement par les équipes de Kellis [Kellis et al., 2004] et Philippsen [Dietrich et al., 2004].

Le projet Génolevures est un excellent cadre d'application réel pour le développement d'une stratégie d'annotation multi-génome. Ce projet bénéficie de résultats expérimentaux et d'analyses bio-informatiques générales (annotation des génomes, visualisation des données) et spécifiques (études thématiques) afin de répondre dans les meilleures conditions possibles à ses objectifs scientifiques. Le travail de cette thèse s'inscrit dans les phases Génolevures 2 et Génolevures 3.

3.4 Conclusion

Comme nous l'avons vu dans ce chapitre, l'annotation d'un génome est un processus long et complexe. Les systèmes d'aide à l'annotation facilitent le travail manuel des annotateurs en automatisant certaines tâches (*e.g.* recherche de séquences homologues dans les banques de données). L'annotation automatique est, de ce fait, plus rapide mais elle donne des résultats de qualité moindre que celle manuelle ou semi-automatique. Or l'annotation génomique doit être la meilleure qualité possible, car elle est le fondement de l'extraction de connaissances pour la compréhension du fonctionnement d'un organisme. De plus, elle guide les analyses expérimentales réalisées ultérieurement.

L'annotation génomique doit ainsi fournir tous les éléments nécessaires au fonctionnement de l'organisme et être cohérente vis-à-vis des contraintes biologiques reconnues par la communauté scientifique. Nous avons vu que les résultats de l'annotation des différents projets n'intègrent pas une vision fonctionnelle globale du génome et ne prennent pas suffisamment en compte son sens biologique. Afin d'accroître la qualité de l'annotation, nous nous proposons d'appliquer ces deux principes en définissant des règles de vérification de la cohérence de l'annotation génomique que nous allons détailler au chapitre 4 et valider au chapitre 5.

Pour autant, seule l'expérimentation biologique permet de vérifier la justesse d'une annotation. Aussi, nous nous sommes intéressés à l'identification des complexes protéiques par la méthode expérimentale de l'électrophorèse en bleu natif et SDS, pour les raisons énoncées dans le chapitre introductif (cf. § 1, p. 7). Nous avons choisi cette méthode d'électrophorèse BN/SDS car elle ne nécessite pas d'étapes préparatives de biologie moléculaire, contrairement aux techniques du double hybride et de purification en tandem. Nous présentons cette partie expérimentale dans le chapitre 6.

TAB. 3.2 – Caractéristiques des levures Génolevures 1.

ass. = état de l'assemblage : partiel (p) ou complet (c),# chr. = nombre de chromosomes, Géno. = taille du génome (Mb), % introns = pourcentage de gènes avec introns, nc = non connu, nd = non disponible.

Sources : Génolevures, sauf pour *C. tropicalis* : BROAD Institute (données au 01/01/2008).

espèce	assem.	# chr.	Géno.	ORF	% introns	% GC	utilisation, particularités
Saccharomyces bayanus var. uvarum	p	16	nc	nc	nc	nc	fermentation alcoolique du vin, Champagne, bière blonde
Saccharomyces exiguus	p	14-16	nc	nc	nc	nc	responsable du goût acide et de l'arôme caractéristique des pains au levain, de certains laits fermentés (kefir)
Saccharomyces servazzii	p	12	nc	nc	nc	nc	fermentation du glucose, galactose
Zygosaccharomyces rouxii	c	7	nd	nd	nd	nd	halotolérant (croissance en milieu salé) et osmorésistant, utilisé pour la fermentation des haricots de soja et de céréales pour la fabrication de la sauce de soja
Lachancea (Saccharomyces) kluyveri	c	8	nd	nd	nd	nd	fermentation alcoolique en anaérobie, production de protéines
Kluyveromyces thermotolerans	c	7	nd	nd	nd	nd	position phylogénétique intermédiaire
Kluyveromyces lactis	c	6	10,6	5327	2,4	38,7	affinage et aromatisation de fromages, industrie laitière, industrie pharmaceutique (production de protéines recombinantes humaines (sérum albumine, composant du plasma sanguin, et interleukine 1β, médiateur de la réponse inflammatoire))
Kluyveromyces marxianus varum marxianus	p	10	nc	nc	nc	nc	fermentation au lactose de lait (kefir, kumys)
Pichia angusta	p	nc	nc	nc	nc	nc	méthylotrophique, utilisé dans l'industrie pharmaceutique (production de protéines recombinantes humaines)
Debaryomyces hansenii varum hansenii	c	7	12,2	6896	4,6	36,3	cryotolérant et halotolérant, activité d'affinage de fromages, croûtage des fromages (protéolyse de la pâte et neutralisation du pH de la croûte), production de vitamine B2
Pichia sorbitophila	p	7	nc	nc	nc	nc	halotolérant
Candida tropicalis	c	12	14,6	6258	0,5	33,1	pathogène humain provoquant des candidoses (infections superficielles de la peau et des muqueuses et infections profondes pulmonaires, urinaires et septicémiques)
Yarrowia lipolytica	c	6	20,5	6437	14,5	49,0	affinage de fromages, fermentation de lait (yaourt), métabolisation d'hydrocarbures et lipides, métabolisme sécrétoire important (la production de protéines recombinantes)

Chapitre 4

Vérification de la cohérence de l'annotation génomique

Sommaire

L'annotation génomique d'un organisme est la base de l'étude de son fonctionnement : elle doit donc être de grande qualité. Comme nous l'avons vu en introduction, l'annotation d'un génome est soumise à des exigences de rapidité et de qualité, la satisfaction de la première pouvant nuire à la seconde et réciproquement. Disposer d'une annotation est déterminant pour mener à bien et au plus tôt les expériences réalisées par la suite. Il est alors indispensable de s'assurer de la bonne qualité d'une annotation, en évaluant celle-ci par la vérification de sa cohérence et de sa complétude.

Je propose une méthode de vérification *in silico* de la cohérence basée sur des règles de logique, méthode peu coûteuse en temps et en argent par rapport à une méthode d'expérimentation biologique. Ces règles s'appliquent sur trois niveaux de façon successive : au niveau du gène, du chromosome et du génome entier. En fait, cet ordre d'application suit naturellement

celui des erreurs par ordre d'importance décroissante. En effet, il faut d'abord définir correctement la base de l'annotation, *i.e.* les éléments identifiés au cours de l'annotation syntaxique, avant de considérer l'annotation du génome dans sa globalité, *i.e.* l'annotation fonctionnelle puis l'annotation relationnelle. De plus, certaines règles sont prévues pour s'appliquer régulièrement au cours de l'annotation d'un génome, afin de prendre en compte au fur et à mesure les nouvelles informations et d'éliminer au plus tôt les erreurs possibles.

Nous décrivons dans un premier temps les objectifs de cette méthode *in silico* puis la stratégie d'application de ces règles. Nous définissons ensuite les domaines de valeurs sur lesquels ces règles s'appliquent, les opérations réalisées sur ces domaines, ainsi que les analyses nécessaires à l'application des règles. Nous définissons enfin les règles de cohérence pour chacun de leurs trois niveaux d'application (gène, chromosome et génome) ainsi que leurs intérêts.

4.1 Objectifs

J'ai développé un ensemble de règles de cohérence avec trois objectifs.

L'objectif prioritaire est de définir un génome qui soit complet et qui ait un sens biologique correct. Le génome d'un organisme contient, sous la forme de séquences ADN, l'ensemble des éléments lui permettant de fonctionner. Il doit donc être considéré comme une entité fonctionnelle, et non pas uniquement comme une liste de ses éléments. Les règles développées s'assurent ainsi que ces éléments obéissent aux contraintes biologiques structurelles. Elles vérifient aussi qu'il ne manque aucun élément nécessaire aux fonctions vitales afin que, d'un point de vue théorique, les éléments définis par l'annotation réalisée pour cet organisme, lui permettent d'être un système fonctionnel. Cela permet de pallier au problème des 'bonnes' CDS écartées du fait de seuils de détection fixés arbitrairement trop hauts pour l'annotation. En recherchant particulièrement les gènes essentiels, les règles de cohérence mettent en évidence ces oublis éventuels. De plus, en atteignant cet objectif, la qualité de l'annotation est améliorée, ce qui *in fine* permet de limiter la propagation d'erreur et de cibler au plus juste les expériences ultérieures.

Un deuxième objectif est de déterminer l'origine d'une différence d'annotation observée entre les génomes d'espèces proches d'un point de vue phylogénique. Soit cette différence est due à une erreur d'annotation (omission ou interprétation erronée) ; dans ce cas, les annotateurs doivent corriger l'annotation. Soit elle révèle une caractéristique de l'espèce considérée, et par conséquent les gènes responsables de la spéciation expliqueraient cette différence. Les règles vont ainsi porter sur la couverture de l'annotation par l'ontologie GO, la conservation ou non des voies métaboliques et des interactions protéine-protéine par comparaison avec *S. cerevisiae*. Ce second objectif a pour pré-requis la satisfaction du premier objectif car l'annotation de ces génomes doit d'abord être cohérente individuellement afin que leur comparaison soit réalisée sur des bases solides.

Le troisième objectif est d'améliorer la gestion du travail des annotateurs et de garantir l'homogénéité de l'annotation réalisée (annotation manuelle ou semi-automatique). Les règles de cohérence sont, à ce niveau, associées à des règles d'annotation. Par exemple, ces règles peuvent vérifier la structure du commentaire ajouté à la prédiction d'un gène et rédigé par

les annotateurs selon les règles d'annotation. L'intégration de ces règles de cohérence à un logiciel d'annotation peut ainsi améliorer l'annotation et les analyses basées sur celle-ci.

Lors d'un projet d'annotation et de comparaison génomiques tel que Génolevures, certaines règles s'appliquent plusieurs fois au cours de l'annotation, telles que les règles vérifiant la structure d'un gène codant une protéine (présence de codon start et stop, respect du cadre de lecture...). Les annotateurs jugent ainsi de l'avancée du travail réalisé et identifient les cas délicats (notamment pour l'identification des exons et intron(s) d'un gène). Les règles portant sur le voisinage des gènes et sur l'ensemble du génome s'appliquent après un premier travail d'annotation syntaxique.

4.2 Stratégie

Il nous semble raisonnable que ces règles de cohérence vérifient l'adéquation de l'annotation avec les lois admises au sein de la communauté scientifique décrivant aussi bien le système biologique fonctionnel représenté par ces éléments (que nous pouvons associer au fond de l'annotation) que la structure des éléments composant l'annotation (que nous pouvons comparer à la forme de l'annotation). Ces règles s'appliquent selon l'approche ascendante ou 'bottom-up' (cf. fig. 4.1) à partir de faits primaires, issus de l'annotation génomique, et de faits résultant d'analyses spécifiques.

Cette stratégie définit ainsi trois classes de règles :

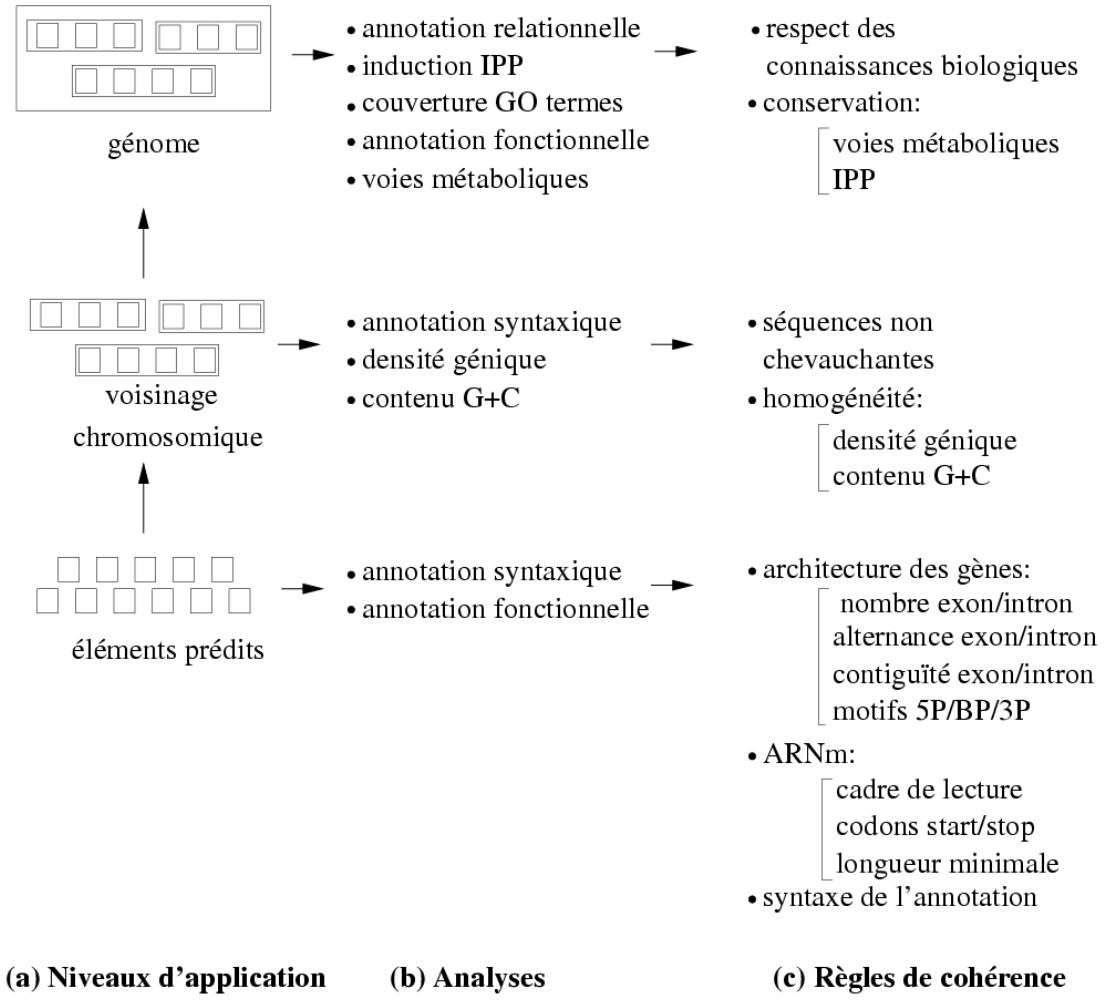
1. les règles qui s'appliquent aux gènes indépendamment les uns des autres : elles améliorent la qualité de l'annotation car elles diminuent le taux d'erreur dans la structure des éléments prédits lors de l'annotation,
2. les règles qui s'appliquent au chromosome : elles améliorent la qualité de l'annotation car elles vérifient la localisation et la répartition des éléments prédits les uns par rapport aux autres sur un chromosome et
3. les règles qui s'appliquent à l'échelle du génome entier : elles améliorent la qualité de l'annotation génomique car elles vérifient l'exhaustivité de l'annotation et le recoupement des connaissances.

J'ai choisi cette approche car l'annotation ne peut être correcte si ses fondements, *i.e.* les gènes, ne le sont pas. Cette approche ascendante suit l'ordre défini par l'importance relative des erreurs présentes dans une annotation de génome pour chaque niveau.

La figure 4.1 représente, selon l'approche ascendante, pour chacun des niveaux d'application des règles de cohérence, les analyses préalables à la définition des règles (lecture de bas en haut et de gauche à droite) :

1. niveau élémentaire :
   - analyses : annotation syntaxique, annotation fonctionnelle
   - règles : architecture de l'ARNpm (nombre d'exons et d'introns, alternance des types exon et intron, contiguïté des bornes introniques et exoniques, respect des motifs introniques), ARNm (codons start et stop, cadre de lecture), longueur minimale de la protéine, syntaxe de l'annotation,

FIG. 4.1 – Application ascendante des règles de cohérence pour l'annotation génomique Pour chaque niveau d'application (a), les analyses spécifiques (b) permettent l'application de règles précises (c).



2. niveau chromosomique :
   - analyses : annotation syntaxique, densité génique, contenu en G+C (cf. § 4.3.2.2 p. 80),
   - règles : non chevauchement des éléments d'ADN génomique d'intérêt, homogénéité de la densité génique pour les différents chromosomes ainsi que du pourcentage G+C,
3. niveau génomique :
   - analyses : annotation fonctionnelle, annotation relationnelle, couverture de la GO, induction des voies métaboliques et interactions protéine-protéine,
   - règles : respect des connaissances biologiques, conservation des voies métaboliques et interactions protéine-protéine.

Ces règles de cohérence ont deux avantages : la vérification de l'intégrité de la base de données d'annotation et la comparaison de génomes.

4.3 Définitions et pré-requis

Les règles développées vérifient la satisfaction des contraintes biologiques principales et essentielles admises et validées par l'ensemble de la communauté scientifique en biologie, et sont basées sur la logique.

Pour décrire les contraintes au moyen de règles, nous avons besoin de nommer les objets qui correspondent aux éléments biologiques manipulés. Nous avons également défini un ensemble d'opérations utilisées pour le formalisme des règles. Ces opérations correspondent en fait aux mécanismes biologiques et aux transitions applicables à ces objets.

4.3.1 Domaines de valeur et opérations

Les domaines de valeurs et opérations sont définis selon une vue de la base de données orientée objet Génolevures dans laquelle est stocké l'ensemble des informations liées à l'annotation génomique des espèces. Cette base de données garantit, grâce à des mécanismes internes non traitées ici, l'intégrité des données selon le modèle imposé (*e.g.* toute séquence contenue dans la base de données doit être de longueur non nulle).

Ces domaines de valeur reprennent ainsi les objets biologiques primordiaux : les chromosomes, les gènes prédits codant une molécule d'ARN (ARNr, ARNt, ARNsn...) ou une protéine, les éléments non-géniques (centromères, rétrotransposons, pseudogènes), les séquences d'ARNpm, les séquences d'ARNm, les séquences protéiques, les exons et les introns. Nous considérons ainsi les objets ou domaines de valeur suivants :

- *Chromosomes* : ensemble des chromosomes, dont chacun se caractérise par un nom, l'organisme auquel il appartient, une séquence biologique ADN simple brin orientée 5'-3' :
  $chromosome = \langle nom, espece, seq \rangle$
- *Regions* : ensemble des régions de chromosomes, dont chacune se caractérise par un nom, un chromosome sur lequel elle est localisée, les coordonnées sur ce chromosome :
  $region = \langle nom, chromosome, debut, fin \rangle$

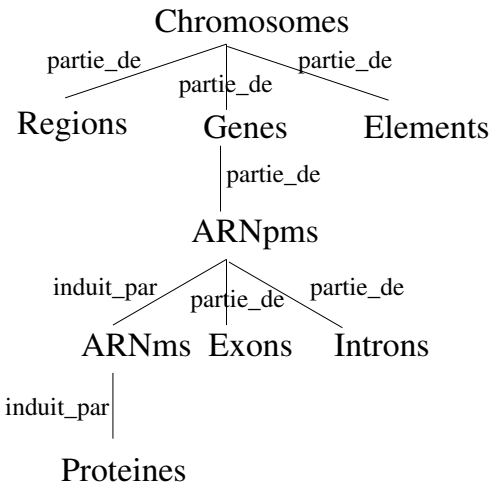
ce domaine a été rajouté pour des raisons de commodité dans la définition des contraintes et des règles,

- *Genes* : ensemble des gènes, dont chacun se caractérise par un nom, une annotation, le chromosome sur lequel il est localisé et sur quel brin, les coordonnées sur ce chromosome : $gene = \langle nom, annotation, chromosome, debut, fin, brin \rangle$
- *Elements* : ensemble des éléments d'ADN identifiés non géniques lors de l'annotation génomique tels que les rétrotransposons, les pseudogènes ou les centromères : $element = \langle nom, annotation, chromosome, debut, fin, brin, type \rangle$
- *ARNpms* : ensemble des molécules d'ARN pré-messager, dont chacune se caractérise par un nom, le gène à partir duquel elle est transcrite, et par une séquence : $ARNpm = \langle nom, gene, seq \rangle$
- *Exons* : ensemble des exons, dont chacun se rattache à une molécule d'ARN pré-messager par un couple de coordonnées, son numéro de position dans la succession d'exons constituant l'ARNpm : $exon = \langle ARNpm, numero, debut, fin \rangle$
- *Introns* : ensemble des introns, dont chacun se rattache à une molécule d'ARN pré-messager par un couple de coordonnées, son numéro de position dans la succession d'introns constituant l'ARNpm : $intron = \langle ARNpm, numero, debut, fin \rangle$
- *ARNms* : ensemble des molécules d'ARN messager (et par conséquent épissée), dont chacune se caractérise par un nom, l'ARNpm à partir de laquelle elle est épissée, et par une séquence : $ARNm = \langle nom, ARNpm, seq \rangle$
- *Proteines* : ensemble de protéines, dont chacune se caractérise par un nom, l'ARNm à partir de laquelle elle est traduite, et par une séquence : $proteine = \langle nom, ARNm, seq \rangle$
- *Motifs* : ensemble des motifs dont chacun correspond à une séquence : $motif = \langle seq \rangle$
  Ces motifs sont regroupés en liste selon leur type :
  - *liste_5P* : liste des motifs spécifiques des sites d'épissage donneurs,
  - *liste_BP* : liste des motifs spécifiques des sites d'épissage accepteurs,
  - *liste_3P* : liste des motifs spécifiques des sites de point de branchement,
  - *liste_start* : liste des motifs spécifiques des codons initiateurs sur l'ARNm, en fait un seul motif : $liste\_start = \{AUG\}$,
  - *liste_stop* : liste des motifs spécifiques des codons terminateurs sur l'ARNm : $liste\_stop = \{UAA, UAG, UGA\}$.

Ces domaines de valeur se rapprochent de ceux décrits dans la SO ("Sequence Ontology") [Eilbeck et al., 2005]. Nous nous restreignons ici aux éléments développés dans nos règles. Les relations entre ces domaines de valeur de niveau inférieur (objet fils) et objets de niveau supérieur (objet parent) sont de type 'induit_par' et 'partie_de' et peuvent être représentées sous forme de structure ascendante arborescente (cf. fig 4.2).

Par exemple, sur la figure 4.2, *Exons* est une partie de *ARNpms*.

FIG. 4.2 – Relation entre domaines de valeur.



L'épissage des introns se faisant au niveau de l'ARNpm pour donner un ARNm, j'ai choisi de rattacher les exons et introns à l'ARNpm plutôt qu'au gène, bien que ce dernier soit à l'origine de la séquence de l'ARNpm. De plus, cela nous permet de nous affranchir de l'orientation du gène sur le brin d'ADN et d'avoir ainsi, pour chaque exon et intron : $debut < fin$.

Les attributs des domaines de valeur, autres que ceux référençant les domaines de valeurs, ont des types précis :

- les coordonnées *debut* et fin sont des entiers naturels (pour tout objet donné : $debut < fin$),
- les séquences *seq* sont du texte : les alphabets de séquence ADN et ARN sont les alphabets à quatre lettres ({ `A`,`T`,`G`,`C`} pour l'ADN, { `A`,`U`,`G`,`C`} pour l'ARN) ; les séquences protéiques sont composées de l'alphabet de vingt lettres (cf. chap1. § 2.13),
- le brin *brin* peut prendre deux valeurs : $+1$ pour le brin sens, -1 pour le brin anti-sens,
- les autres attributs sont de type texte.

L'accession à un attribut d'un objet est notée *objet.attribut*. Quelques exemples :

- *gene.seq* : accède à la séquence de *gene*.
- *gene.debut* : accède à la coordonnée inférieure de *gene*,
- *exon.numero* : accède au numéro de l'exon.

Nous définissons ensuite les opérations appliquées sur ces domaines de valeur :

- *coder* : $Genes \rightarrow Proteines$: obtention à partir d'un gène de la protéine,
- *transcrire* : $Genes \rightarrow ARNpms$: transcription du gène en ARN pré-messager,
- *episser* : $ARNpms \rightarrow ARNms$: épissage des introns sur l'ARNpm pour obtenir l'ARNm,
- *traduire* : $ARNms \rightarrow Proteines$: traduction de l'ARNm en protéine,

- *exprimer* : $Genes \rightarrow Proteines$: obtention de la séquence protéique à partir de la séquence ADN du gène,
- *code* : $Genes \rightarrow ARNms$: séquence codante d'un gène,
- *subseq* : $Chromosomes \rightarrow Chromosomes$: obtention d'une sous-séquence à partir d'une séquence d'un objet de la coordonnée x à la coordonnée y comprise : $subseq(objet.seq, x, y)$,
- *nombre* : $Chromosomes \rightarrow Chromosomes$: obtention du nombre d'unité de bases x (acide nucléique ou aminé) présent dans la séquence d'un objet : $nombre(objet.seq, x)$,
- *homologie* : $Genes \rightarrow Genes$: relation d'homologie entre deux gènes X et Y, et notée par $homologie(X, Y)$,
- *orthologie* : $Genes \rightarrow Genes$: relation d'orthologie entre deux gènes X et Y, et notée par $orthologie(X, Y)$,
- *interaction* : $Proteines \rightarrow Proteines$: relation d'interaction entre deux protéines X et Y notée par $interaction(X, Y)$,
- *gene_de* : $Regions \rightarrow Genes$: ensemble de gènes appartenant à une région *region* d'une molécule,
- *longueur* : $Chromosomes \rightarrow Chromosomes$: obtention de la longueur de la séquence d'un objet,
- *seq* : $Chromosomes \rightarrow Chromosomes$: obtention de la séquence d'un objet.

D'après le diagramme commutatif ci-dessous représentant certains domaines de valeurs et opérations :

$$\begin{array}{ccc} gene & \xrightarrow{coder} & proteine \\ {\scriptstyle seq}\downarrow & & \downarrow{\scriptstyle seq} \\ sequence\ ADN & \xrightarrow[exprimer]{} & sequence\ aa \end{array}$$

nous pouvons noter que :

$$\begin{aligned} \forall\ gene \in Genes,\ seq(coder(gene)) &= exprimer(seq(gene)) \\ &= traduire(episser(transcrire(seq(gene)))) \end{aligned}$$

Certaines analyses et règles requièrent des paramètres qui peuvent être fixés par les annotateurs :

- la longueur minimale d'une protéine,
- les distances entre les trois sites d'épissage sur l'intron,
- la longueur de séquence considérée pour le calcul de la densité génique,
- le pourcentage de variation d'une valeur autorisée.

4.3.2 Analyses basées sur les faits

L'application des règles nécessite diverses analyses préliminaires, spécifiques de leur niveau d'application. Une partie de ces analyses mettent en évidence les faits primaires, stockés dans la base de données Génolevures, issus directement de l'annotation génomique de l'espèce.

L'autre partie des analyses met en évidence des faits secondaires, résultant de l'analyse de l'annotation syntaxique et fonctionnelle ou faisant appel à l'annotation relationnelle.

4.3.2.1 Analyses basées sur les faits primaires

Les analyses préliminaires basées sur les faits primaires sont des analyses descriptives, basées sur l'annotation syntaxique et l'annotation fonctionnelle telles que nous l'avons définie au chapitre précédent.

Analyse syntaxique L'analyse syntaxique (cf. § 3.1.3 p. 43) identifie plusieurs caractéristiques :

- les gènes codant une protéine, avec la définition du codon initiateur et du codon terminateur, la présence des exons et intron(s) potentiels, la CDS associée et la séquence protéique résultante,
- les gènes codant une molécule d'ARN,
- les autres éléments non codants tels que les centromères, les rétrotransposons, les pseudogènes.

L'analyse syntaxique est utilisée pour le développement de règles de cohérence appliquées aux niveaux de l'élément et du voisinage chromosomique.

Analyse fonctionnelle L'annotation fonctionnelle (cf. § 3.1.5 p. 52) est l'attribution du commentaire décrivant le rôle de chaque élément identifié. Si l'élément du génome nouvellement annoté est homologue à un élément déjà annoté, alors ce commentaire est inféré (cf. § 3.1.4) à partir de celui de l'élément déjà connu. L'analyse fonctionnelle est utilisée pour le développement des règles appliquées à trois niveaux : élémentaire, chromosomique et génomique.

4.3.2.2 Analyses basées sur les faits secondaires

Une partie des règles de cohérence nécessite une analyse des informations primaires issues de l'annotation syntaxique et de l'annotation fonctionnelle. Ces règles concernent la vérification de la cohérence de l'annotation au niveau du voisinage des gènes et au niveau du génome.

Les analyses requises portent sur l'étude de la densité génique, le contenu en G+C. Ces analyses se basent aussi sur l'annotation fonctionnelle et l'annotation relationnelle de l'organisme nouvellement séquencé, ainsi que sur les connaissances disponibles pour l'organisme de référence le plus proche de cet organisme. La quantité de données disponibles pour l'organisme de référence (*S. cerevisiae* dans le cadre de notre travail) permet d'inférer des prédictions de même nature pour les génomes nouvellement séquencés. Ainsi, nous pouvons utiliser des analyses portant sur l'inférence d'interactions protéine-protéine, de voies métaboliques, d'annotation génomique en termes ontologiques (utilisation de la GO).

Densité génique La densité génique est le nombre de gènes présents sur une région d'ADN génomique *region*. Elle est spécifique de chacun des organismes.

$$densite(region) = \frac{|gene\_de(region)|}{longueur(region)} \tag{4.1}$$

La densité génique s'exprime en fait pour un gène : par exemple 1 gène/2 kpb, 1 gène/10 kpb.

Contenu en G+C Les régions codantes ont un pourcentage en nucléotides `G` et `C` supérieur aux régions non codantes [Guigó and Fickett, 1995, Oliver and Marín, 1996]. Le contenu en G+C est le nombre de nucléotides `G` et `C` dans une séquence donnée *region* :

$$contenu\_GC(region) = \frac{nombre(seq(region), G) + nombre(seq(region), C)}{longueur(region)} \tag{4.2}$$

Des études chez divers organismes [Payseur and Nachman, 2002] dont *S. cerevisiae* [Dujon et al., 1994] ont montré que la densité génique et le contenu en G+C étaient positivement corrélés. La densité génique et le contenu en G+C interviennent dans les règles de cohérence qui s'appliquent au niveau du chromosome et du génome.

Inférence d'annotation GO Un génome nouvellement séquencé peut être annoté avec l'ontologie GO, par inférence des annotations GO d'organismes apparentés. Dans le cadre du projet Génolevures, l'annotation GO d'un génome de levure nouvellement séquencé, est inférée à partir de celle de l'organisme de référence *S. cerevisiae*. Puis la couverture GO de ce génome est comparée à celle de *S. cerevisiae*. Comme nous l'avons vu, les sur-représentations ou sous-représentations d'éléments pour certaines catégories fonctionnelles par rapport à la couverture GO de *S. cerevisiae*, révèlent alors soit une divergence phylogénique, soit une erreur d'annotation. Une recherche bibliographique ou une expérience biologique peut confirmer ou non les propriétés de la nouvelle espèce. Le développement d'une GO spécifique de la levure par le consortium GO, en version complète ('GO yeast') et en version minimale ('GO slim yeast') nous sert de référence pour la comparaison avec les nouvelles levures annotées lors du projet Génolevures.

Soit $go\_terme(g_A)$ le terme GO associé au gène g_A de l'organisme A, et g_B le gène prédit de l'espèce B en cours d'annotation et homologue à g_A :

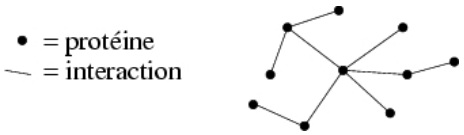
$$go\_terme(g_B) = go\_terme(g_A) \quad si\ homologie(g_A, g_B) \tag{4.3}$$

$$\begin{array}{l} Soit\ GO_{g_A} = \bigcup_{g_A} go\_terme(g_A), \\ GO_{g_B} = \bigcup_{g_A} go\_terme(g_A) \quad si\ homologie(g_A, g_B) \end{array} \tag{4.4}$$

L'homologie est établie par comparaison significative entre les deux séquences et appartenance à la même famille protéique.

Inférence d'interactions protéine-protéine La relation d'orthologie entre protéines permet la prédiction d'IPP. L'orthologie est la relation d'homologie existante entre deux protéines appartenant chacune à un organisme différent.

L'ensemble des IPP d'un organisme forme un réseau qui peut être vu comme un graphe dont les nœuds sont les protéines et les arêtes les interactions :



Ce réseau est donc formé de couples de protéines en interaction.

Soit P_i^A la ième protéine de A alors l'ensemble $IPP(A)$ des couples de protéines de l'organisme A en interaction est défini par :

$$IPP(A) = \{(P_i^A, P_j^A) | \forall\ i,\ j,\ i \neq j,\ interaction(P_i^A, P_j^A)\} \tag{4.5}$$

D'après les relations d'orthologie entre les protéines des organismes A et B, un sous-ensemble $part\_IPP(B)$ de $IPP(B)$, $IPP(B)$ étant l'ensemble des couples de protéines en interaction de B, peut être extrapolé :

$$\begin{aligned} part\_IPP(B) \quad = \quad & \{(P_i^B, P_j^B) | \forall\ i,\ j,\ i \neq j, \\ & orthologie(P_i^A, P_i^B) \wedge\ orthologie(P_j^A, P_j^B) \\ & \wedge\ interaction(P_i^A, P_j^A)\} \end{aligned} \tag{4.6}$$

$part\_IPP(B)$ est ici un sous-ensemble de $IPP(B)$ car il existe diverses méthodes prédictives permettent d'inférer des interactions protéine-protéine parmi lesquelles les méthodes se basant sur les événements de fusion de gènes ou bien le profil phylogénique des protéines (cf. § 3.1.5.2 p.53). Chacune de ces méthodes permet de compléter l'ensemble des IPP pour un organisme. Ces méthodes pourraient être utilisées ultérieurement.

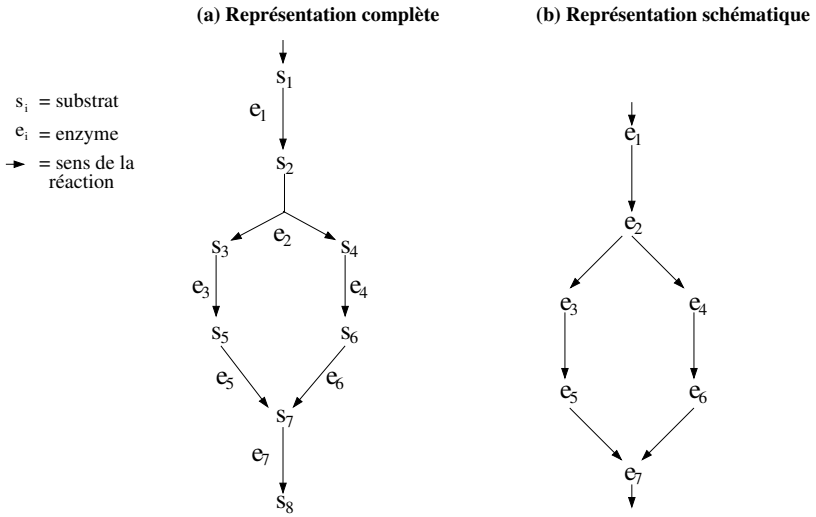
Inférence de voies métaboliques La reconstruction des voies métaboliques d'un organisme inconnu par inférence de celles connues chez l'espèce modèle *S. cerevisiae* permet d'identifier des gènes manquants ou sur-représentés pour ces voies. Alors que les fonctions absentes peuvent être dues à des erreurs d'annotation, la sur-représentation peut traduire le métabolisme et les processus physiologiques particuliers de l'organisme nouvellement séquencé.

Une voie métabolique est une succession de réactions métaboliques catalysées par des protéines enzymatiques à partir d'au moins un substrat initial pour former au moins un produit final. Elle peut être considérée comme un ensemble ordonné et alterné d'enzymes et de substrats, avec au moins un substrat en entrée et au moins un substrat en sortie.

Soit la voie métabolique $VM(A)$ (cf. fig. 4.3 (a)) chez l'organisme A, avec $Substrats$ l'ensemble des substrats s_i^A présents dans $VM(A)$, et $Enzymes$ l'ensemble des enzymes e_j^A catalysant les réactions chimiques. $VM(A)$ se définit par l'ensemble de tous les couples

FIG. 4.3 – Représentation d'une voie métabolique
Une voie métabolique peut se représenter de deux façons : (a) représentation complète avec l'ensemble de ses substrats et enzymes ; (b) représentation schématique avec seulement les enzymes.

zz



d'éléments enzyme-substrat en interaction, couples ordonnés selon l'orientation de la réaction :

$$\begin{aligned} VM(A) \;=\; & \{(s_i^A, e_k^A), (e_k^A, s_j^A) |\; \forall\; i,\; j,\; k, \\ & s_i^A \in Substrats \wedge\; s_j^A \in Substrats \wedge\; \{s_i^A\} \cap \{s_j^A\} \\ & \wedge\; e_k^A \in Enzymes\} \end{aligned} \tag{4.7}$$

Cet ordonnancement des enzymes et substrats nous permet de relier plusieurs substrats impliqués dans une réaction chimique catalysée par une enzyme, ainsi que plusieurs substrats issus (ce sont alors des produits) d'une réaction chimique catalysée par une enzyme.

Les relations d'orthologie entre les organismes A et B permettent ainsi d'induire, comme pour les IPP, une partie $part\_VM(B)$ des voies métaboliques de l'organisme B à partir de celles connues chez A :

$$\begin{aligned} part\_VM(B) \;=\; & \{(s_i^B, e_k^B), (e_k^B, s_j^B) |\; \forall\; i,\; j,\; k, \\ & orthologie(s_i^A, s_i^B) \wedge\; orthologie(s_j^A, s_j^B) \wedge\; orthologie(e_k^A, e_k^B) \\ & \wedge\; (s_i^A, e_k^A) \in VM(A) \wedge\; (e_k^A, s_j^A) \in VM(A)\} \end{aligned} \tag{4.8}$$

Une voie métabolique est fonctionnelle lorsque ses enzymes nécessaires à son parcours de part en part et le substrat initial sont présents. Par simplification, nous faisons abstraction des substrats (le substrat initial provenant d'un apport extérieur à la cellule, les suivants issus des réactions enzymatiques) pour nous intéresser uniquement aux enzymes. La voie métabolique peut alors être représentée par un graphe orienté, dont les nœuds sont les enzymes et les arcs les réactions (cf. fig. 4.3 (b)).

Par exemple, la voie métabolique v représentée fig. 4.3 se compose des enzymes e_1, e_2, e_3, e_4, e_5, e_6 et e_7.

Cette voie métabolique V possède deux chemins possibles $chemin_1$ et $chemin_2$ que nous représentons par :

$$\begin{aligned} chemin_1(V) = &\quad (e_1, e_2, e_3, e_5, e_7) \\ chemin_2(V) = &\quad (e_1, e_2, e_4, e_6, e_7) \end{aligned}$$

De façon plus générale, une voie métabolique VM peut ainsi être décomposée en l'ensemble $\mathcal{C}$ de ses chemins possibles :

$$\mathcal{C}_{VM} = \{chemin_i | \ \forall e \in chemin_i, \ e \in VM\} \tag{4.9}$$

Une voie métabolique fonctionnelle se traduit alors par la présence des enzymes permettant de la traverser de part en part.

L'inférence de voies métaboliques faisant appel aux relations de similarités entre les protéines, nous utilisons les relations établies par les familles de protéines définies par le projet Génolevures. En effet, ces relations se basent sur l'analyse de la séquence protéique, élément fixe, contrairement à l'annotation fonctinnelle qui peut évoluer.

Ces analyses, basées sur des faits primaires et secondaires, représentent les informations issues de l'annotation d'un génome. Les règles de cohérence, appliquées à ces informations, permettent ainsi de vérifier la qualité de l'annotation génomique réalisée.

4.4 Règles de cohérence

Les règles de cohérence que j'ai développées s'appliquent à trois niveaux différents : élémentaire, chromosomique et génomique. Cet ordre est étroitement lié au processus d'annotation qui consiste à identifier d'abord les éléments intéressants sur la molécule d'ADN génomique puis leur attribuer une ou plusieurs fonctions et finalement établir leurs relations mutuelles.

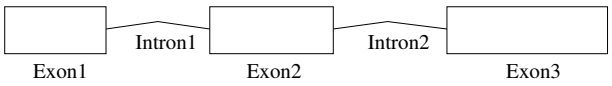
Ces règles vérifient d'une part les contraintes biologiques telles que l'architecture des gènes à intron(s), l'expression d'un gène en protéine, d'autre part les contraintes émises par les responsables du projet d'annotation telles que la syntaxe du commentaire. Elles vérifient également les contraintes biologiques spécifiques de l'espèce considérée telle que les motifs d'épissage ou la longueur minimale d'une protéine.

4.4.1 Règles élémentaires

Les règles de cohérence appliquées au niveau élémentaire concernent les gènes. Elles vérifient l'architecture des gènes à intron et la justesse de la définition de l'ARNm issu de ces gènes, ainsi que la syntaxe de l'annotation.

4.4.1.1 Architecture de l'ARNpm

Nous considérons ici seulement les ARNpm ayant une architecture décomposée en exons et intron(s).
Pour rappel, un ARNpm, composé de 3 exons et 2 introns, peut être représenté ainsi :



La structure correcte d'un ARNpm, et par conséquent du gène à partir duquel il est transcrit, se vérifie à trois niveaux :
- le nombre d'éléments de type exon et intron,
- l'alternance des types exoniques et introniques et
- la contiguïté des éléments.

Nombre d'exon et d'intron Un ARNpm $ARNpm$ a sa structure en exon et intron complète s'il compte un exon de plus que d'intron(s) :

Règle 1 $\forall i,\ j,\ k,\ |exon_i.ARNpm_k| = |intron_j.ARNpm_k| + 1$

L'absence de détection d'exon 1 et d'intron 1 ne sera pas signalée par cette règle. Cette situation se produit quand les motifs d'épissage aux sites 5' et 3' n'ont pas une séquence consensus ou au contraire, il existe plusieurs choix possibles de motifs d'épissage.

Alternance exon / intron Un ARNpm $ARNpm$ est une suite alternée de séquences exoniques *exon* et intronique(s) *intron* telle que les première et dernière séquences sont toujours un exon (cf. fig ci-dessus). Les exons et introns appartenant à un même ARNpm doivent vérifier les égalités suivantes, avec $i > 0$ et $j > 0$:

Règle 2 *Soit* $exon_i$ *et* $intron_j$ *tels que* $\forall\ i,\ j,\ exon_i.ARNpm = intron_j.ARNpm$*, ils vérifient les égalités suivantes :*
- $exon_i.numero = intron_j.numero$
- $exon_{i+1}.numero = intron_j.numero + 1$ *et*
- *soit* $Numeros$ *l'ensemble des numero,* $1 \in Numero$*.*

L'absence d'exon 1 et d'intron 1 sera détectée par cette règle (cf. § 4.4.1.1). La non satisfaction de cette règle souligne en fait la difficulté de détection de l'exon 1 par rapport aux exons suivants (cf. § 3.1.3.1 p. 45).

Les ARNpm qui ont échoué à la règle 1 (pour cause d'absence d'exon 1) échouerons également à cette règle pour ce qui concerne l'alternance exon 1/intron 1.

Contiguïté des bornes exoniques et introniques De plus, les exons et introns d'un ARNpm sont correctement définis s'ils sont contigus :

Règle 3 *Soit* $exon_i$ *et* $intron_j$ *tels que* $\forall\ i,\ j,\ exon_i.ARNpm = intron_j.ARNpm$*,, ils vérifient les égalités suivantes :*

- *jonction exon - intron :*
  - $exon_i.numero = intron_j.numero$,
  - $exon_i.fin = intron_j.debut - 1$,
- *jonction intron - exon :*
  - $exon_{i+1}.numero = intron_j.numero + 1$,
  - $exon_{i+1}.debut = intron_j.fin + 1$.

Les ARNpm sans exon 1 vont échouer pour la vérification de la jonction exon 1-intron 1. Dans le cadre du système d'annotation mis en place pour le projet Génolevures, les autres cas d'erreur devraient être dus à des erreurs humaines de "copier-coller" des séquences exoniques ou introniques lors de la définition de ces éléments sur la plate-forme d'annotation.

Le respect de l'alternance exon/intron et de la contiguïté des bornes exoniques et introniques est important pour la suite car la séquence protéique prédite est issue de l'épissage des introns tels qu'ils sont définis dans la base de données. ;

Motifs introniques Un intron se caractérise par trois sites distincts de motifs particuliers :

1. le site donneur en 5'de l'intron, d'une taille de 6 bases,
2. le point de branchement au sein de la séquence intronique, d'une taille de 7 bases et
3. le site accepteur en 3' de l'intron, d'une taille de 3 bases.

Les motifs de ces sites sont bien conservés parmi les espèces (cf. § 3.1.3.1), aussi doivent-ils correspondre à un profil défini à partir des motifs trouvés de façon expérimentale.
Lorsqu'un gène ou ARNpm annoté a un ou plusieurs introns, il est aisé d'obtenir les motifs 5' et 3', contrairement au motif du point de branchement dont la position n'est pas conservée dans la base de donnée.

Un élément intronique prédit *intron*, appartenant à un ARNpm $ARNpm$, est valide si et seulement si ses motifs donneur d et accepteur a appartiennent respectivement aux listes de motifs acceptés $liste\_5P$ et $liste\_3P$ et qu'au moins un des motifs de $liste\_BP$ soit présent dans la séquence intronique amputée des sites donneur et accepteur :

Règle 4 *L'intron est correctement prédit si :*
- $subseq(seq(ARNpm), intron.debut, intron.debut + 5) \in liste\_5P$,
- $subseq(seq(ARNpm), intron.fin - 2, intron.fin) \in liste\_3P$,
- $subseq(seq(ARNpm), intron.debut + 6, intron.fin - 3) \cap\ liste\_BP \neq \varnothing$.

Les erreurs capturées par cette règle mettent en évidence des motifs d'épissage non consensuels. Ceci peut représenter un excès de prédiction d'introns dans les séquences géniques et provoquer ainsi l'augmentation de faux positifs parmi les séquences codantes prédites.

4.4.1.2 Validité de l'ARNm

Les règles de cohérence appliquée à l'ARNm (séquence ARN épissée) portent sur le respect des contraintes pour une bonne traduction de l'ARNm en protéine.

Codons start et stop Les triplets de nucléotides aux extrémités de tout ARNm $ARNm$ codant une protéine, doivent correspondre chacun à un motif contenu dans les listes $liste\_start$ pour le codon initiateur) et $liste\_stop$ pour le codon terminateur :

Règle 5 *L'ARNm est correctement prédit si :*
- $subseq(seq(ARNm), 1, 3) \in liste\_start$,
- $subseq(seq(ARNm), longueur(ARNm) - 2, longueur(ARNm)) \in liste\_stop$.

L'absence de codon initiateur peut révéler une séquence sans exon 1, tout comme l'absence de codon terminateur pour une séquence sans exon final.

Cadre de lecture La séquence d'un ARNm doit respecter la traduction de cette séquence nucléique en protéine selon le premier cadre de lecture (ce cadre débute au premier nucléotide de la séquence). La longueur de l'ARNm doit être un multiple de 3. Sa lecture selon le code génétique ne doit pas faire apparaître de codon stop avant celui de la fin de la séquence en 3'. Soit *proteine* la protéine obtenue après transcription du gène *gene* en ARN, épissage des introns éventuels et traduction, sa longueur de séquence doit être égale au tiers de celle de son ARNm $ARNm$:

Règle 6 $\forall\ i, longueur(proteine_i) = 3 \times longueur(ARNm_i)$ *avec* $proteine_i = traduire(episser(transcrire(gene_i)))$,

Les erreurs détectées concernent principalement les ARNpm sans exon 1, car le cadre de lecture n'aura qu'une chance sur trois d'être dans la bonne phase de lecture. Les ARNpm ayant les bornes exoniques et introniques non contiguës et non corrigées échoueront également à cette règle.

Longueur minimale de séquence codante Les annotateurs fixent arbitrairement la longueur minimale de détection d'une séquence codante une protéine, pour l'organisme nouvellement séquencé :

Règle 7 *Une protéine peut être prédite si :* $longueur(ARNm) \geqslant seuil$

Ce seuil peut être fixé d'après des observations réalisées chez une espèce apparentée. Dans le cadre de Génolevures, les experts biologistes se sont basés sur les observations chez *S. cerevisiae* dont les deux plus petites séquences protéiques ont une taille de 24 aa, soit une longueur d'ARNm de 72 nt. Les experts de Génolevures ont ainsi fixé le seuil à 60 nt.

Cette règle permet de diminuer le nombre de modèles de gènes proposés à l'annotateur. Mais cela permet surtout de diminuer le nombre de faux positifs. En effet, plus le seuil minimal est bas, plus il y a de chance d'obtenir une séquence codante par hasard. Les responsables de projet d'annotation préfèrent ainsi manquer les rares cas réels plutôt que d'avoir une sur-prédiction de séquences codantes. Ces cas réels non détectés pourront être recherchés ultérieurement *in silico* (cf.§ 4.4.3 p. 91) ou *in vivo* (par exemple : par l'étude d'EST, cf. § 3.2.1.3 p.59).

4.4.1.3 Syntaxe de l'annotation

La syntaxe de l'annotation, en tant que commentaire attribué à un gène, est également soumise à une vérification. Cette syntaxe fait partie d'un ensemble de règles d'annotation dont l'élaboration est réalisée par le consortium Génolevures et dirigée par Pascal Durrens. Ces règles permettent une homogénéité de la rédaction du commentaire.

Le respect de cette syntaxe est vérifié d'après ces règles d'annotations. La rédaction de l'annotation d'un modèle de gène suit deux cas :

1. soit le modèle de gène est un gène connu de l'espèce : la description contient dans l'ordre : l'identifiant UniProt, l'espèce, le nom du gène, et la fonction si elle est connue ; tout modèle de gène dont la description ne commence pas par un niveau de similarité, est un gène connu et doit donc vérifier la syntaxe requise pour ce cas,
2. soit le modèle de gène a, ou non, un niveau de similarité avec un gène annoté : la description commence par un des termes appartenant au vocabulaire contrôlé ('highly similar to', 'similar to', 'weakly similar to', 'some similarities with', 'conserved hypothetical protein', 'no similarity').

Si la description commence par l'un des termes du vocabulaire contrôlé, trois possibilités d'annotation se présentent :

1. le modèle de gène est similaire à un gène de *S. cerevisiae* : la description doit contenir dans l'ordre : le niveau de similarité ; l'identifiant UniProt sinon celui de SGD ; '*Saccharomyces cerevisiae*' ; le nom systématique du gène ; le nom usuel du gène ; la fonction si elle est connue,
2. le modèle de gène est similaire à un gène d'une espèce autre que *S. cerevisiae* : la description doit contenir dans l'ordre : le niveau de similarité ; l'identifiant UniProt ; le nom de l'espèce ; le nom systématique du gène ; le nom usuel éventuel du gène ; la fonction si elle est connue,
3. le modèle de gène ne ressemble à aucun gène connu de façon convaincante : la description est définie par 'no similarity'.

En cas de non respect de cette syntaxe, certaines corrections peuvent être automatisées par l'utilisation d'une grammaire, telles les erreurs orthographiques du vocabulaire contrôlé, les absences d'identifiant et de nom de gènes (s'ils existent).

Par ailleurs, les annotations par homologie de séquence sont reportées à partir de données issues de bases de données à un instant t. Or ces données évoluent au fil du temps, suite à des analyses expérimentales ou bio-informatiques, et leur existence peut ainsi être confirmée ou non. Étant effectuée plusieurs mois après cet instant t, la validation de l'annotation doit tenir compte de l'évolution du contenu des bases de données. L'existence du gène qui conditionne l'annotation du modèle de gène, est confirmée par sa présence dans les bases de données au moment de l'application des règles de cohérence. Ceci évite la propagation d'erreurs dans les bases de données en reprenant un gène qui n'a plus lieu d'être.

Par exemple, l'annotation d'un modèle de gène par rapport à *S. cerevisiae* suit la règle suivante :

Règle syntaxique 1 *L'annotation Annotation commençant par un niveau de similarité et contenant les mots ressemblants à "*Saccharomyces cerevisiae*" doit débuter par les groupes de mots ordonnés suivants :*

Annotation ="^ <niveau_similarite> <identifiant> <Saccharomyces cerevisiae> *<nom_systematique>"*

Une des règles de production est alors :

Règle de production syntaxique 1 *Si Annotation contient* <Saccharomyces cerevisiae> *alors <identifiant> est placé juste avant* <Saccharomyces cerevisiae>*,*

si Annotation contient <Saccharomyces cerevisiae> *alors <nom_systematique> est placé juste après* <Saccharomyces cerevisiae> *,*

L'application de cette règle permet d'avoir un commentaire ayant une structure et un vocabulaire homogène. Par ailleurs, cette règle permet de détecter les fautes de syntaxe (dans la mesure où celle-ci est précisée), les informations obsolètes (*e.g.* les gènes supprimés d'une base de donnée, identifiant erroné). Cette règle permet ainsi, dans la mesure du possible, d'éviter la propagation des erreurs.

L'ensemble de ces règles élémentaires peut être appliqué régulièrement au cours de l'annotation d'un génome. Ces règles signalent ainsi les modèles de gènes complexes qui méritent plus d'attention de la part des annotateurs. De plus, ces règles révèlent l'état d'avancement de l'annotation. En effet, tout modèle de gène non validé échouera sur plusieurs règles : absence d'annotation, pas de codon start, longueur de la séquence protéique non proportionnelle à la séquence génique...

4.4.2 Règles chromosomiques

Ces règles s'applique aux éléments, géniques et non géniques, appartenant à une même région chromosomique, voire au chromosome entier. Elles vérifient les deux contraintes biologiques communément admises suivantes : l'absence de chevauchement entre ces éléments localisés sur un même chromosome et la présence d'un seul centromère par chromosome. Elles vérifient également une contrainte observée au moins chez les levures : l'homogénéité des chromosomes selon leur pourcentage en G+C et en densité génique.

4.4.2.1 Séquences non chevauchantes

Chez les eucaryotes, les séquences d'intérêt biologique ne se chevauchent pas, même si elles sont en anti-sens les unes par rapport aux autres, sauf dans de très rares exceptions (ARNsn localisé dans un intron, gènes de rétrotransposon). Par exemple le gène snRN14 codant un ARNsn (impliqué dans le complexe de l'épissage) se trouve dans l'intron du gène YER007c chez *S. cerevisiae*.

L'ensemble $\mathcal{E}$ des éléments *Genes* et *Elements* est trié par ordre croissant de leur coordonnée *debut* :

$$\forall\ e \in \{Genes, Elements\}, \mathcal{E} = \{e_i | \forall\ i,\ j\ \in \{1, \ldots, n\}\ :\ e_i < e_j \Leftrightarrow e_i.debut < e_j.debut\} \tag{4.10}$$

Soit deux éléments e_i et e_{i+1} en position respective i et $i+1$.

Règle 8 *Deux éléments e_i et e_{i+1} ne chevauchent pas si :*
$\forall\ i \in \{1, \ldots, n-1\},\ e_i.fin < e_{i+1}.debut.$

Le cas de séquence codante présentant de l'épissage alternatif n'intervient pas ici. En effet, dans ce cas, c'est un seul et même gène qui code un seul ARN pré-messager qui, lui, sera épissé de plusieurs façon, donnant ainsi autant d'ARNm et donc de protéines.

Quoi qu'il en soit, les cas de chevauchement détectés doivent être soumis à une expertise humaine afin d'être corrigés ou validés. Cette vérification permet également de s'assurer de l'unicité de chacun des éléments annotés.

4.4.2.2 Présence de centromère

Le centromère est la zone du chromosome qui assure une bonne séparation des chromosomes lors de la division cellulaire. Chaque chromosome n'a qu'un seul et unique centromère.

Règle 9 $\forall\ i,\ j,\ element_i.chromosome = chromosome_j.nom \wedge\ element_i.type = \text{'centromere'},$
$|element_i.nom| == 1$

Cette règle n'est pas vérifiée pour les chromosomes dont l'annotation omet le centromère. En effet, lors de l'annotation d'un génome, les annotateurs se répartissent le travail selon leurs compétences. Les éléments particuliers tels que les retrotransposons, les centromères, les éléments transposables sont ainsi recherchés en dehors de la recherche des séquences codantes (fonctionnelles ou devenues non fonctionnelles telles que les pseudogènes) par un nombre très réduit de personnes (le plus souvent une seule). Il serait alors surprenant (mais pas impossible), dans ces conditions, qu'une personne (ou deux) attribue deux centromères à un chromosome.

4.4.2.3 Homogénéité de l'annotation des chromosomes

La densité génique et le contenu en G+C sont deux caractéristiques propres à chaque organisme. Cependant, les résultats sont différents selon les organismes considérés. Chez les levures, la densité génique et le contenu en G+C reflètent l'annotation syntaxique [Dujon et al., 1994]. Ces deux observations suivent cependant une variation cyclique sur l'ensemble des chromosomes et sont corrélées positivement.

Mais ce n'est pas le cas pour l'ensemble des eucaryotes. Par exemple, chez l'homme, l'annotation révèle une densité génique de 23 gènes/Mpb pour le chromosome 19, et une densité de 6 gènes/Mpb pour le chromosome 4 [Venter et al., 2001]. Aussi, les règles concernant la densité génique et le contenu en G+C doivent être appliquées selon l'organisme ou la branche phylogénique considérée.

Densité génique Mises à part les régions télomériques et subtélomériques des chromosomes (régions situées aux extrémités du chromosome et contenant peu ou pas de gènes), un génome de levure devrait avoir une répartition homogène de ses gènes. L'annotateur peut ainsi fixer un taux de variation δ accepté pour une densité génique attendue $densite_{attendue}$ pour un chromosome :

Règle 10 $|densite(region) - densite_{attendue}| \leqslant \delta$

Une variation trop importante de cette densité génique, fixée par l'expérimentateur, le long d'un chromosome révèlerait soit une zone particulière en terme de composition, soit des oublis lors de l'annotation syntaxique.

La densité génique peut être observée au niveau du chromosome entier, d'une région de taille donnée (par exemple 100 000 Mb) ou du génome entier. Une forte densité génique au niveau chromosomique révèlerait une sur-prédiction de gènes pour ce chromosome, ou à l'opposé, une faible densité génique révèlerait une faible prédiction de gènes. Mais une faible densité génique peut aussi bien résulter de la composition de l'ADN génomique à cet endroit-là : présence de longues séquences intergéniques (ADN "poubelle"), présence de séquences répétées non codantes... La zone impliquée nécessiterait alors une seconde annotation syntaxique.

Contenu en G+C Selon les connaissances actuelles [Guigó and Fickett, 1995, Oliver and Marín, 1996], le contenu en G+C (calculé en pourcentage) est plus important dans les séquences codantes. Chez les levures, ce contenue varie significativement d'une espèce à l'autre [Dujon et al., 2004]. Aussi, après annotation syntaxique, les séquences prédites codantes devraient suivre cette règle :

Règle 11 *Soit* $Codant = \{Genes\}$,
$contenu\_GC(Codant) > contenu\_GC(\overline{Codant})$

Le contenu en G+C étant spécifique d'une espèce, l'annotateur peut fixer un taux δ de variation accepté pour le contenu en G+C attendu $contenu\_GC_{attendu}(Codant)$ pour les séquences codant la région considérée, d'après les observations chez une espèce apparentée :

Règle 12 *Soit* $Codant = \{Genes | Genes \in region\}$,
$|contenu\_GC(Codant) - contenu\_GC_{attendu}(Codant)| \leqslant \delta$

Cette règle peut être appliquée sur chacun des chromosomes comme sur l'ensemble des chromosomes.

Cette règle nous permet de détecter la sur-prédiction (ou la sous-prédiction) de séquences codantes dans le cas d'un contenu en G+C supérieur (ou inférieur) au contenu toléré.

Par ailleurs, chez les levures, le contenu en G+C a été démontré comme étant corrélé positivement avec la densité génique [Dujon et al., 1994] pour une région considérée :

Règle 13 $\frac{densite(region)}{contenu\_GC(region)} > 0$

La vérification de non chevauchement des séquences peut être appliquée régulièrement au cours de l'annotation. Les règles liées à la densité génique et au contenu en G+C doivent être appliquées après un premier tour d'annotation syntaxique.

4.4.3 Règles génomiques

Les règles appliquées à l'échelle du génome s'intéressent principalement à l'ensemble des gènes codant les protéines de l'organisme ayant une annotation fonctionnelle.

Certaines règles vérifient la prise en compte des connaissances biologiques actuelles sur l'organisme annoté et les organismes qui lui sont proches. D'autres comparent également l'annotation de l'organisme nouvellement séquencé à celle d'autres organismes, afin d'en souligner les différences, conséquences de la spéciation d'une espèce ou d'erreurs d'annotation. L'application de ces règles orientée génomique comparée est pertinente seulement si les annotations des génomes considérés sont correctes. Ces règles vérifient ainsi la complétude de l'annotation.

4.4.3.1 Intégration de connaissances biologiques

L'annotation d'un génome doit tenir compte des connaissances biologiques disponibles lors de sa réalisation. Ces connaissances peuvent être spécifiques de l'organisme si celui-ci a déjà fait l'objet d'études expérimentales, ou bien communes aux espèces dont il est proche d'un point de vue phylogénique.

Ainsi le génome annoté doit contenir les éléments (protéines, ARNr, ARNt) décrits dans la littérature scientifiques et dans les bases de données généralistes et spécialistes. Par exemple, *Y. lipolytica* est connue pour son activité de dégradation des hydrocarbures et lipides [Müller et al., 1998, Pignede et al., 2000] : l'annotation de son génome doit avoir identifié au moins une des enzymes lipases caractérisées auparavant de façon expérimentale. L'annotation prend en compte les connaissances à un instant t, et d'après les règles d'annotation, un modèle de gène identifié comme gène connu et déjà publié (répertorié par un nom et un identifiant dans la banque de données UniProt), ce modèle de gène prend le nom et l'identifiant de ce gène.

Soit $\mathcal{C} = \{c_1, \ldots, c_n\}$ l'ensemble des n éléments connus c de l'organisme A en cours d'annotation, $\mathcal{C}$ doit être présent dans l'ensemble des éléments identifiés au cours de l'annotation de cet organisme :

Règle 14 *Soit $\mathcal{X}$ l'ensemble des Genes et Elements de A identifiées au cours de l'annotation, et $\mathcal{C}$ l'ensemble des éléments connus avant annotation, alors $\mathcal{C} \subset \mathcal{X}$*

L'annotation du nouveau génome doit identifier les éléments lui permettant d'assurer ses fonctions nécessaires. Ces fonctions sont communes aux organismes vivants car elles sont soumises à une très forte pression de sélection : la moindre défection entraîne la mort cellulaire.

L'ensemble de ces fonctions cellulaires assure en théorie le fonctionnement d'une cellule minimale. La définition de cette cellule minimale fait l'objet de nombreuses recherches [Smalley et al., 2003, Gabaldon et al., 2007] dans les domaines de la biologie et de la modélisation des systèmes ('Systems Biology'). Ces recherches se concentrent sur l'étude d'organismes procaryotes, plus faciles d'étude que les eucaryotes grâce à leur génome plus simple (absence de gène fragmenté, génome plus dense (jusqu'à 80% d'ADN codant) et plus petit (1 à 5.10^6pb contre $1{,}5.10^7$ à 90.10^{12}pb).

Les techniques de génie génétique (manipulation de l'ADN génomique entraînant une modification de la séquence par délétion, insertion ou changement de la séquence ADN) [Sambrook and Russel, 2001] permettent d'étudier l'action du gène ciblé par observation de la différence de comportement de la cellule en présence et absence de l'expression de ce gène.

La mutation d'un gène peut ainsi entraîner la mort cellulaire. Le gène muté est alors appelé gène létal car il est essentiel, vital pour la survie de la cellule. Nous considérons comme gènes essentiels les gènes létaux et les gènes synthétiques létaux. Un gène synthétique létal est un gène dont la mutation n'est létale que si un autre gène est muté. L'annotation d'un nouveau génome doit en principe contenir les gènes essentiels mis en évidence pour les organismes apparentés tels que *S. cerevisiae* (cf. p. 107).

Soit $\mathcal{E} = \{e_1, \dots e_n\}$ l'ensemble des n gènes nécessaires pour la branche phylogénique à laquelle appartient l'organisme A considéré, $\mathcal{E}$ doit être présent dans l'ensemble des gènes identifiés au cours de l'annotation de cet organisme :

Règle 15 *Soit $\mathcal{Y}$ l'ensemble des Genes de A et $\mathcal{E}$ l'ensemble des gènes nécessaires pour la branche phylogénique à laquelle appartient A alors $\mathcal{E} \subset \mathcal{Y}$*

Ces deux règles vérifient que l'annotation du génome est cohérente dans le sens où, d'une part, elle intègre les connaissances biologiques relatives à l'espèce considérée, et d'autre part, elle définit un organisme vivant d'un point de vue fonctionnel et physiologique.

4.4.3.2 Conservation des interactions protéine-protéine

La composition des complexes protéiques est bien conservée au cours de l'évolution car soumise à une forte pression de sélection. Chez *S. cerevisiae*, des expériences à grande échelle [Uetz et al., 2000, Ito et al., 2001, Ho et al., 2002] ont permis d'identifier de nombreux complexes protéiques.

Les complexes protéiques de génomes nouvellement annotés de levure peuvent ainsi être inférés à partir de ceux de *S. cerevisiae* et par comparaison entre les gènes des deux espèces. Le complexe protéique $Cplx(A)$ présent chez l'organisme A peut être présent chez l'organisme B si celui-ci possède au moins un homologue pour chacune des protéines p_i de $Cplx(A)$, et par conséquent les gènes, impliquées dans ce complexe.

Règle 16 *Soit $Cplx(A)$ le complexe protéique présent chez l'organisme A tel que*
$Cplx(A) = \{p_1^A, \dots, p_i^A\}$,
le complexe protéique $Cplx(B)$ est conservé chez l'organisme B si
$Cplx(B) = \{p_i^B | \ \forall i, \exists j$ *avec* $homologie(p_j^A, p_i^B)\}$.

Les protéines et complexes nécessaires au fonctionnement de la cellule, tels que ceux responsables de la respiration cellulaire, devraient être conservés. Une protéine nécessaire absente, impliquée ou non dans un complexe, doit alors être recherché de façon plus ciblée que lors de l'annotation syntaxique.

L'annotateur doit s'affranchir des critères de recherche fixés lors de l'annotation syntaxique, tels que la longueur minimale de détection d'une séquence codante. Les méthodes de recherche de similarité de séquences doivent être modifiées. L'annotateur ne recherche plus une homologie de séquence entre la séquence codante prédite et les séquences présentes dans les banques de données. Il recherche une homologie de séquence entre la séquence de la protéine nécessaire, absente jusqu'alors de l'annotation de l'espèce nouvellement séquencée, et la séquence d'ADN génomique de l'espèce considérée.

4.4.3.3 Conservation des voies métaboliques

Nous considérons une voie métabolique comme un graphe orienté dont les nœuds sont les enzymes et les arêtes les réactions.

Une des voies métaboliques d'un organisme nouvellement séquencé est conservée si les enzymes permettant au moins un des parcours du graphe représentant cette voie, sont présentes. Soit $\mathcal{C}_{VM(A)}$ l'ensemble des chemins possibles de la voie métabolique $VM(A)$ de l'organisme de référence A, la voie métabolique $VM(B)$ est conservée chez l'organisme B si les enzymes d'au moins un chemin de $VM(A)$ sont présentes dans l'ensemble des gènes de B.

Règle 17 *La voie métabolique $VM(B)$ est conservée chez l'organisme B si $chemin_x \subset Genes(B)$ tel que $\forall\ x,\ \exists\ chemin_x \in \mathcal{C}_{VM(A)}$.*

Certaines voies métaboliques sont essentielles pour le fonctionnement de la cellule, telle que la voie de la glycolyse. L'absence d'une de ces voies métaboliques chez l'espèce nouvellement séquencée doit entraîner la recherche active d'un ou des gènes responsables des fonctions manquantes.

Par ailleurs, l'analyse des familles de protéines impliquées dans les voies métaboliques peut mettre en évidence les particularités des espèces considérées. Par exemple, *Y. lipolytica* présente une expansion de la famille de protéines à activité enzymatique de lipase. Cela ne signifie pas forcément que toutes ces protéines interviennent dans la même voie métabolique. Ces lipases peuvent être exprimées à différents états de la cellule (pendant les différentes phases du cycle cellulaire), sous diverses conditions (par exemple : nutriments disponibles, stress environnemental...), être localisées dans divers compartiments cellulaires (cytoplasme, lysosomes, vacuole...).

4.5 Conclusion et perspectives

J'ai développé cet ensemble de règles de logique afin de vérifier la qualité de l'annotation génomique d'un organisme nouvellement séquencé. Ces règles de cohérence détectent les séquences incomplètes en terme de structure (absence de codon start ou de codon stop, cadre de lecture non respecté, motif non consensuel...) mais ne les corrigent pas. Ne connaissant pas tous les mécanismes biologiques ni les cas particuliers, nous ne pouvons rejeter de façon automatique un cas qui ne suit pas à la lettre les principes biologiques connus à l'heure actuelle.

La détection d'erreurs (structure incomplète d'un gène, fonction essentielle absente de l'annotation) permet également aux annotateurs de repérer les cas difficiles et de concentrer leurs efforts pour tenter de résoudre ces problèmes. En cela, ces règles aident les annotateurs dans la gestion de leur travail.

De plus, l'analyse des variations entre les génomes permet d'évaluer la prédiction informatique pour l'annotation des génomes et la divergence phylogénique (spécificité fonctionnelle d'un organisme par rapport à un autre).

Ces règles font appel à des règles de logique, ainsi que des faits issus d'analyses classiques dans l'études des génomes (prédiction d'interaction protéine-protéine, densité génique...). Leur mise en œuvre nécessite cependant l'existence de données fiable sur lesquelles se baser et jouant ainsi le rôle de standard. Le projet Génolevures offre un formidable cadre d'application pour ces règles. D'une part, les levures étudiées dans le projet sont suffisamment proches d'un point de vue fonctionnel pour que les différences mises en évidence soient dues à la spéciation. D'autre part, nous disposons d'un organisme de référence pour les levures, *S. cerevisiae*, dont l'annotation génomique est parmi les plus complètes chez les eucaryotes à l'heure actuelle.

Les règles de cohérence définies ici sont développées dans le cadre de l'annotation de génomes eucaryotes de levures. Il se dégage alors plusieurs perspectives de ce travail. Tout d'abord, les organismes pour lesquels ces règles ont été développées, présentent une structure génomique et un fonctionnement biologique généraux. Ainsi, ces règles peuvent être appliquées à d'autres organismes, eucaryotes ou procaryotes. Néanmoins, certaines règles nécessitent d'être adaptées en fonction de l'organisme considéré. Par exemple, le génome procaryote est circulaire, possédant non pas un centromère mais plusieurs origines de réplication. Certaines espèces, eucaryotes et procaryotes, présentent un code génétique différent de celui universel. Par exemple, *Candida albicans* présente les codons initiateurs `ATA`, `ATC` et `ATT`, et le codon terminateur `AGA`. Autres cas particuliers, le génome des microspories et des microalgues rouges ou vertes, organismes unicellulaires eucaryotes, possèdent de nombreux cas de gènes chevauchant.

Ensuite, certaines règles peuvent être définies par d'autres méthodes complémentaires ou validées dans le cadre de certaines espèces. Par exemple, l'utilisation des COG[1] ("Clusters of Orthologous Groups") [Tatusov et al., 1997, Tatusov et al., 2003] permettrait d'établir des relations d'orthologie pour d'autres organismes, étant donné que les familles de protéines utilisées ici concernent seulement les levures du projet Génolevures. Les relations COG pourraient également être utilisées comme méthode complémentaire à celle des familles de protéines de Génolevures. De même, il existe cependant des méthodes plus élaborées pour la prédiction et l'inférence de fonction moléculaire telles que SIFTER [Engelhardt et al., 2005] qui utilise les relations phylogéniques chez les bactéries. Là aussi, ces méthodes pourraient être appliquées aux levures.

[1]Banque de données COG : http ://www.ncbi.nlm.nih.gov/COG/.

Chapitre 5

Mise en œuvre des règles de cohérence

Sommaire

Les règles de cohérence développées dans le chapitre précédent permettent de mettre au jour les imperfections de l'annotation de génome. Nous avons appliqué une partie de ces règles sur deux jeux de données : l'un annoté de façon manuelle lors du projet Génolevures 2 (*C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*), l'autre en cours d'annotation semi-automatique lors du projet Génolevures 3 (*Z. rouxii*, *S. kluyveri* et *K. thermotolerans*).

Dans un premier temps, nous présentons le système d'annotation mis en place pour l'annotation manuelle des levures lors du projet Génolevures 2. Cette plate-forme d'annotation a évolué en système d'annotation semi-automatique, dénommé MAGUS, pour l'annotation des levures du projet G2nolevures 3. Nous présentons ensuite les données biologiques, bibliographiques et bio-informatiques disponibles pour l'application de ces règles. Nous voyons ensuite

la mise en œuvre d'une partie de ces règles et leur application sur les données de Génolevures. Nous analysons les résultats de leur application aux génomes annotés lors du projet Génolevures.

5.1 Système d'annotation pour Génolevures

Comme nous l'avons vu dans le § 3.3 p.66, le projet Génolevures 2 consistait en l'annotation manuelle des levures *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*. Le système d'annotation mis en place pour le consortium est le fruit d'un travail collaboratif avec David Sherman. Ce système devait répondre aux attentes des membres du consortium. La première attente était la mise en place d'un système d'annotation manuelle intégrant les analyses bio-informatiques demandées par les annotateurs afin d'obtenir une annotation de qualité et homogène. La seconde était de pouvoir commencer l'annotation avant la fin de l'assemblage des séquences génomiques afin de disposer de plus de temps pour les analyses nécessaires.

5.1.1 Base de données Génolevures

Les membres de Génolevures portent un grand intérêt à la diffusion de leurs résultats et informations à la communauté scientifique. Cette diffusion se voulait être plus qu'une simple consultation des résultats dans des banques de données généralistes. Dès sa première phase, Génolevures a produit de grandes quantités de données, offrant ainsi une diversité de requêtes et d'analyses pour une ou plusieurs espèces de levure. Aussi le consortium investit depuis 2000, en termes financier et humain, dans la création et l'administration d'une banque de données appelée Génolevures consultable par internet (http ://cbi.labri.fr/Genolevures) et mise à disposition de la communauté scientifique.

Cette banque de données est le fruit d'une étroite collaboration entre les biologistes et les bio-informaticiens du projet de façon à répondre au mieux aux requêtes des biologistes, premiers utilisateurs du site Génolevures.

David Sherman et moi avons maintenu et amélioré la base de données existante dans laquelle sont stockées les données Génolevures 1. Cette base de données relationnelle PostgreSQL est interrogée via une interface graphique (http ://cbi.labri.fr/Genolevures/Genolevures.php) basée sur un ensemble de scripts réalisés en Php. Les requêtes peuvent se faire par mot-clef, nom de gène, expression rationnelle. Les données peuvent être téléchargées sous différents formats : FASTA, EMBL ou XML.

Les résultats de Génolevures 2 constituent à ce jour les principales données du site Génolevures. Celui-ci présente également des études thématiques réalisées par des membres du consortium telles que la conservation des voies métaboliques [Iragne, 2006], Génosplicing [Neuvéglise, 2005] ou YETI ("Yeast Transport Information") [De Hertogh et al., 2002], banque de donnée dédiée aux transporteurs membranaires chez *S. cerevisiae* et dont j'ai réalisé la partie ingénierie [De Hertog et al., 2003].

5.1.2 Système d'annotation manuelle

Le système d'annotation mis en place pour l'annotation manuelle des levures *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*, se base sur le logiciel CAAT-Box. Sa facilité de paramétrage et la structure arborescente des données utilisées (sous forme de fichiers et de répertoires) nous ont permis d'enrichir et de personnaliser ses fonctionnalités selon les besoins des biologistes et les particularités des génomes étudiés.

5.1.2.1 Système CAAT-Box

Le système CAAT-Box ("Contig-Assembly Annotation and Tool-Box") [Frangeul et al., 2004] a été développé par Lionel Frangeul, de l'Institut Pasteur, pour l'annotation de génomes bactériens. CAAT-Box regroupe plusieurs modules qui permettent, entre autres, de commencer l'annotation des ORF prédits pendant la phase de finition de l'assemblage et de présenter et modifier ces annotations via une interface web. Lionel Frangeul a amélioré CAAT-Box pour l'adapter aux nouveaux besoins des annotateurs de Génolevures et a également corrigé quelques bugs que nous avons rencontrés lors de la phase de préparation des données pour l'annotation.

D'un point de vue pratique, un premier module de CAAT-Box détecte tous les ORF (phase ouverte de lecture, délimitée par deux codons terminateurs) de longueur fixée par l'annotateur et leur attribue un identifiant. À chaque ORF correspond une fiche individuelle de protéine ou IPF, *i.e.* un fichier plat contenant toutes les informations relatives à cette IPF, ayant comme nom l'identifiant de l'ORF ainsi qu'un numéro de version. Les informations liées à une IPF sont son nom, sa localisation sur un chromosome, la phase de lecture, la séquence protéique potentielle, l'annotation donnée par l'expert, mais également un commentaire éventuel (utile pour les cas complexes d'annotation), les IPF voisins, et les liens vers les résultats d'analyses complémentaires telles que la recherche d'alignements de séquences (*e.g.* de type BLAST), la prédiction du potentiel codant (par GeneMark), le profil d'hydrophobicité (par Toppred). Ainsi, toute nouvelle analyse peut être facilement intégrée dans une IPF grâce à l'insertion du lien pointant vers le ficher résultat. CAAT-Box fait automatiquement la prédiction des gènes codant les protéines. La prédiction des gènes codant les molécules d'ARN, des transposons et rétrotransposons a fait l'objet d'une annotation à part par les experts biologistes spécialisés dans leurs études ainsi que d'une création des fiches IPF correspondantes de façon semi-automatique.

Le module de visualisation de CAAT-Box centralise ensuite sur une seule page toutes les informations disponibles pour cette IPF afin d'aider les annotateurs dans leur travail. Les règles d'annotation émises par les membres du consortium (cf. § 4.4.1.3 p. 87) garantissent une homogénéité de la syntaxe et du niveau informatif du commentaire attribué à l'élément prédit. Par ailleurs, la prise en compte du voisinage proche de l'IPF permet de passer d'une IPF à une autre de proche en proche.

Lors d'un passage d'une version d'assemblage à une autre, un module permet la transposition des informations pour chaque IPF sur la nouvelle IPF, de même identifiant mais ayant un numéro de version différent. Par défaut, le module de visualisation ne tient compte que

des IPF ayant la version la plus récente.

Seuil minimal de détection d'ORF Le paramètre le plus important pour l'annotation est en fait celui fixant la longueur minimale de détection d'un ORF. Les annotateurs ont fixé ce seuil minimal à 60 nt, d'après les observations faites chez *S. cerevisiae*. Le nombre d'ORF potentiels, et par conséquent d'IPF à annoter, obtenu avec ce paramètre était acceptable pour *C. glabrata*, *D. hansenii* et *K. lactis*(environ 10 000 IPF obtenues) mais non pour *Y. lipolytica* (environ 30 000 IPF). En effet, la taille du génome de *Y. lipolytica* (∼20 Mpb) étant presque le double des autres, *S. cerevisiae* compris, le nombre de petits ORF prédits augmentait considérablement. Aussi nous avons relevé le seuil à 80 nt. Ce choix est en fait un compromis entre le nombre d'IPF à annoter et le nombre de faux positifs générés qui pourraient être validés par les annotateurs. Ainsi l'annotation de *Y. lipolytica* portait sur la validation d'environ 13 500 IPF. Ce nombre n'est pas très éloigné de ceux pour les trois autres espèces car si *Y. lipolytica* présente une taille génomique double par rapport aux autres levures (*C. glabrata*, *D. hansenii*, *K. lactis* et *S. cerevisiae*), il s'agit d'une extension de l'ADN intergénique. Lorsqu'une première annotation fût faite, nous recherchâmes les petits ORF potentiels uniquement dans les régions intergéniques définies par la validation des CDS prédites par les annotateurs. Ainsi 61 ORF potentiels d'une taille comprise entre 60 et 80 nt furent finalement annotés.

Même s'il était fonctionnel, les fonctionnalités d'analyse et de visualisation des données du système CAAT-Box devaient être améliorées. En effet, CAAT-box ne permet pas la détection des introns, présents chez les organismes eucaryotes tels que les levures. Par ailleurs, il fallait tirer parti du fait de disposer de quatre génomes de levures apparentées. En l'occurrence, il était intéressant pour l'annotateur de visualiser aisément, pour une IPF d'un génome, la présence de séquences homologues chez les autres levures, qu'il s'agisse de *S. cerevisiae* ou bien des trois autres levures en cours d'annotation.

5.1.2.2 Améliorations du système d'annotation

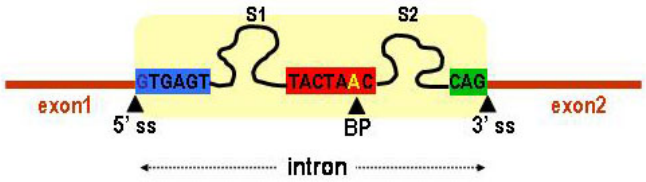
Les améliorations de l'existant portent sur la détection *ab initio* d'introns potentiels et la visualisation des données disponibles ainsi que la navigation d'un génome à l'autre. Pour chacune de ces améliorations, nous présentons d'abord le contexte et les souhaits des membres du consortium puis l'amélioration en elle-même.

Recherche d'intron Le logiciel CAAT-Box a été développé à l'origine pour l'annotation de génomes bactériens. Or, chez les procaryotes, l'architecture des gènes codant une molécule d'ARN ou une protéine est plus simple que celle des gènes eucaryotes qui peuvent contenir un ou plusieurs introns. De plus, l'annotation du génome de *S. cerevisiae* avait révélé que 3,8% des gènes étaient prédits avec un ou plusieurs introns. Des analyses précédentes réalisées, entre autres, par des membres du consortium [Bon et al., 2003, Neuvéglise, 2005], avaient mis en évidence un certain nombre de motifs d'épissage pour les site 5', 3' et le point de branchement. Ces analyses permettaient aussi d'estimer une gamme de distances pour les distances entre le

site 5' et le point de branchement d'une part, et le point de branchement et le site 3' d'autre part (distances désignées respectivement S1 et S2 sur la fig. 5.1).

FIG. 5.1 – Structure d'un intron chez les levures hémiascomycètes.

Les séquences nucléotidiques indiquées correspondent aux motifs consensus des sites d'épissage en 5' (5'ss), au point de branchement (BP) et en 3' (3'ss) chez *S. cerevisiae* (source : [Neuvéglise, 2005]).



Les analyses présentées aux annotateurs devaient ainsi intégrer la prédiction d'introns pour les IPF d'après ces critères. Nous avons choisi de développer une méthode combinatoire de prédiction d'intron basée sur la présence des motifs au site 5', point de branchement et au site 3', pour les distances entre ces sites donnés par les observations précédentes. Il y avait à cela plusieurs raisons. Tout d'abord, nous disposions des motifs précis ainsi que des gammes de distances S1 et S2. Ensuite, les experts souhaitaient que tous les introns répondant à ces critères soient détectés.

Les introns potentiels trouvés étaient ensuite attribués aux IPF. Il pouvait y avoir plusieurs introns potentiels pour une même IPF. La séquence codante était ensuite recalculée en fonction de chacun de ces introns et du nouveau cadre de lecture. Cette potentielle séquence épissée était ensuite comparée aux séquences présentes dans les banques de données par alignement de type BLAST afin de voir si le résultat de l'épissage choisi était probant ou non.

Les algorithmes 1 et 2 retracent notre stratégie de détection des introns, présentée ici pour la recherche sur le brin '+'. Les données d'entrées sont les motifs pour les sites d'épissage en 5', au point de branchement et en 3', les distances S1 et S2, et les chromosomes (ou les contigs) d'une espèce de levure à annoter. Les données de sortie sont la liste des introns exprimés par leur position sur un chromosome (ou contig). L'algorithme 3 retrace l'assignation de chacun des introns aux IPF. Seule la méthode pour le brin '+' est représentée. Les données d'entrée sont les introns trouvés précédemment ainsi que la position et l'orientation des IPF sur le brin de la molécule d'ADN.

Les annotateurs pouvaient visualiser, entre autres, l'ensemble des introns prédits pour les IPF grâce au GBrowse, un outil de visualisation d'annotation de génome développé par GMOD ("Generic Model Organism Database Toolkit", http ://www.gmod.org) (cf. ci-après).

Annotation orientée génomique comparée Le projet Génolevures 2 consistait, dans un premier temps, à annoter les quatre levures simultanément. Les levures étant proches d'un point de vue phylogénique, les membres du consortium souhaitaient réaliser une annotation orientée génomique comparée. En annotant une IPF pour un génome, l'annotateur devait

Algorithme 1 trouver la liste des introns sur le brin '+'.

Entrées: la liste des couples nom et séquence des chromosomes $\mathcal{C}$, la liste des motifs 5P $\mathcal{M}_{5P}$, la liste des motifs BP $\mathcal{M}_{BP}$,la liste des motifs 3P $\mathcal{M}_{3P}$,les distances S1 et S2, le sens sens.

Sorties: la liste des introns $\mathcal{I}$

1: $\mathcal{I} \leftarrow ()$
2: $\mathcal{M}_{BP3P} \leftarrow ()$ {la liste des motifs BP3P}
3: **pour tout** bp dans $\mathcal{M}_{BP}$ **faire**
4: **pour tout** $3p$ dans $\mathcal{M}_{3P}$ **faire**
5: **pour** $s = 0$ à $S2$ **faire**
6: Ajouter le motif $bp.\{s\}3p$ à $\mathcal{M}_{BP3P}$
7: **fin pour**
8: **fin pour**
9: **fin pour**
10: $table\_5P \leftarrow ()$ {le tableau des couples de motifs 5P et de leur coordonnée sur les chromosomes}
11: $table\_BP3P \leftarrow ()$ {le tableau des couples de motifs BP3P et de leur coordonnée sur les chromosomes}

avoir accès facilement aux prédictions et informations pour la région homologue pour les autres génomes.

La visualisation des IPF et de leurs résultats d'alignement est réalisée par la mise en place du GBrowse. Cet outil de visualisation de génome permet la représentation d'une séquence biologique avec ses caractéristiques. Le système permet de se déplacer sur la séquence biologique ou de faire afficher les informations. Dans le cadre de l'annotation réalisée par Génolevures, la séquence biologique est le contig ou le chromosome d'une espèce (suivant l'état d'avancement de l'assemblage des séquences). Les caractéristiques pouvaient être par exemple les modèles de gènes à valider intégrant la prédiction des introns, ou bien les séquences homologues, d'après un alignement de type BLAST, provenant des autres levures. L'annotateur peut choisir les informations qu'il souhaite visualiser en sélectionnant les pistes désirées (*e.g.* les gènes codant les ARNr, les séquences d'une espèce particulière homologues à la région d'intérêt).

Par ailleurs, en fin d'annotation, les annotateurs ont pu bénéficier d'une classification des IPF, mêlant les IPF validés et ceux encore non annotés, en famille de protéines grâce à une analyse réalisée par Hélène Ferry et Macha Nikolski [Dujon et al., 2004, Nikolski and Sherman, 2007].

L'annotation des levures *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica* fut ainsi réalisée manuellement par une soixantaine d'annotateurs. Le résultat de ce travail communautaire est mis à disposition sur le site Génolevures. Il a également fait l'objet d'une publication [Dujon et al., 2004]. Les travaux bio-informatiques que nous avons développés ont également fait l'objet de publications [Sherman et al., 2004, Sherman et al., 2006]. Le fait de prendre part nous-même à l'annotation présentait l'avantage de corriger rapidement les problèmes rencontrés et d'apporter les améliorations, notamment en ce qui concernait la prédiction des introns et la visualisation des données.

Algorithme 2 trouver la liste des introns sur le brin '+' (suite).

pour tout (nom, c) dans $\mathcal{C}$ **faire**
  $k \leftarrow 0$
  $d \leftarrow 0$
  tantque $k \leq longueur(\mathcal{M}_{5P})$ **faire**
    $5p \leftarrow \mathcal{M}_{5P}[k]$
    $coord \leftarrow$ coordonnee$(5p, c, d)$ {Recherche du motif 5p dans la séquence c à partir de la position d}
    si $coord = -1$ **alors**
      $k \leftarrow k + 1$
      $d \leftarrow 0$
    sinon
      Ajouter $(coord, 5p)$ dans $table\_5P$
      $d \leftarrow coord + 1$
    fin si
  fin tantque
  $k \leftarrow 0$
  $d \leftarrow 0$
  tantque $k \leq$ longueur$(\mathcal{M}_{BP3P})$ **faire**
    $bp3p \leftarrow \mathcal{M}_{BP3P}[k]$
    $coord \leftarrow$ coordonnee$(bp3p, c, d)$ {Recherche du motif bp3p dans la séquence c à partir de la position d}
    si $coord = -1$ **alors**
      $k \leftarrow k + 1$
      $d \leftarrow 0$
    sinon
      Ajouter $(coord, bp3p)$ dans $table\_BP3P$
      $d \leftarrow coord + 1$
    fin si
  fin tantque
  Trier $table\_5P$ par coordonnées croissantes
  Trier $table\_BP3P$ par coordonnées croissantes
  $i \leftarrow 0$
  tantque $i \leq$ longueur$(table_5P) - 1$ **faire**
    $(coord5P, 5p) \leftarrow table\_5P[i]$
    $j \leftarrow 0$
    tantque $j \leq$ longueur$(table_BP3P) - 1$ **faire**
      $(coordBP3P, bp3p) \leftarrow table\_BP3P[j]$
      si $coordBP3P < coord5P$ **alors**
        $j \leftarrow j + 1$
      sinon si $coordBP3P - coord5P > S1$ **alors**
        $j \leftarrow$ longueur$(table_BP3P)$
      sinon
        Ajouter $(nom, coord5P, coordBP3P + longueur(bp3p) - 1, sens)$ dans $\mathcal{I}$ {c'est le bon intervalle}
        $j \leftarrow j + 1$
      fin si
    fin tantque
    $i \leftarrow i + 1$
  fin tantque
fin pour
Retourner $\mathcal{I}$

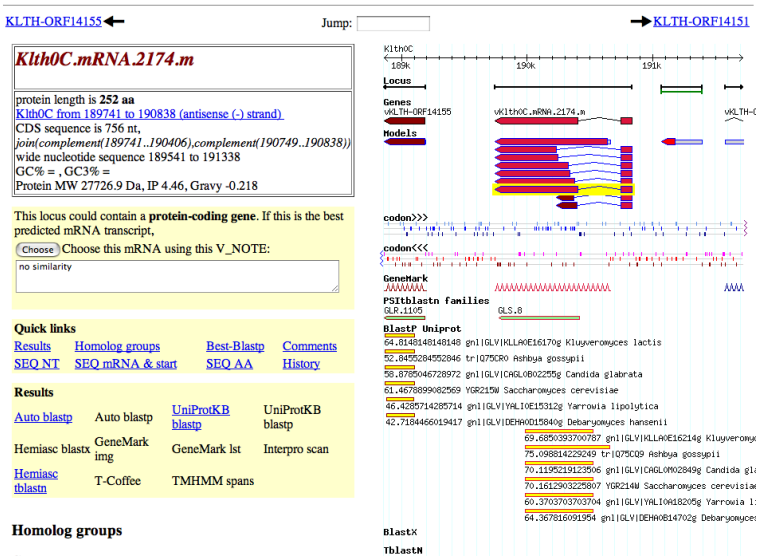
Algorithme 3 assigner les introns à des IPF sur le brin '+'.

Entrées: La liste des introns (nom chromosome, debut,fin et sens) $\mathcal{I}$, la liste des IPF (nom chromosome, debut, fin et sens) $\mathcal{F}$

Sorties: La liste des couples (intron, IPF) $\mathcal{A}$

$\mathcal{A} \leftarrow ()$
Trier $\mathcal{I}$ par chromosome et coordonnées croissantes
Trier $\mathcal{F}$ par chromosome et coordonnées croissantes
$chromosome\_courant \leftarrow nul$
pour tout ipf dans $\mathcal{F}$ **faire**
  $(chromosomeF, debutF, finF, sensF) \leftarrow ipf$
  si $chromosomeF \neq chromosome\_courant$ **alors**
    $chromosome\_courant \leftarrow chromosomeF$
    $fin\_IPF\_precedent \leftarrow -1$
  fin si
  $i \leftarrow 0$
  tantque $i \leq$ longueur($\mathcal{I}$) **faire**
    $intron \leftarrow element(\mathcal{I}, i)$
    $(chromosomeI, debutI, finI, sensI) \leftarrow intron$
    si $chromosomeI = chromosomeF$ **alors**
      si $sens =' +'$ **alors**
        si $finI < debutF - 1$ {l'intron n'est pas juxtaposé en 5'} **alors**
          $i \leftarrow i + 1$
        sinon si $debutI > finF + 1$ {l'intron n'est pas juxtaposé en 3'} **alors**
          $i \leftarrow$ longueur($\mathcal{I}$)
        sinon si $debutI < fin\_IPF\_precedent$ { l'intron est dans l'IP précédente} **alors**
          $i \leftarrow i + 1$
        sinon
          Ajouter $(intron, ipf)$ dans $\mathcal{A}$
          $i \leftarrow i + 1$
          $fin\_IPF\_precedent \leftarrow finF$
        fin si
      fin si
    sinon si $chromosomeI < chromosomeF$ **alors**
      $i \leftarrow i + 1$
    sinon
      $i \leftarrow$ longueur($\mathcal{I}$)
    fin si
  fin tantque
fin pour
Retourner $\mathcal{A}$

FIG. 5.2 – Page d'annotation d'un modèle de gène par le système MAGUS.
Le modèle de gène, parmi ceux proposés par le système MAGUS, dont l'annotateur regarde les informations est surligné en jaune.



5.1.3 Système d'annotation semi-automatique

Pour le projet Génolevures 3, les membres du consortium ont souhaité conserver l'aspect annotation manuelle orientée génomique comparée. Ainsi les levures *Z. rouxii*, *S. kluyveri* et *K. thermotolerans* sont annotées simultanément en utilisant une nouvelle plate-forme d'annotation appelée MAGUS, développée par l'équipe INRIA Magnome (Models and Algorithms for the Genome), dirigée par David Sherman et basée au LaBRI. Le système MAGUS (Magnome genome understanding system) [1] est un outil collaboratif pour l'analyse comparative et l'annotation des génomes complets apparentés, comme c'est le cas pour le projet Génolevures 3. Comme le système précédent basé sur CAAT-Box, il met à disposition des annotateurs, grâce à un navigateur internet, les séquences génomiques et leurs caractéristiques, les analyses *in silico* (*e.g.* prédictions d'ORF, alignements de type BLAST, GeneMark...). L'ensemble de ces caractéristiques et analyses est visualisé grâce au GBrowse et aux pages comportant les informations élémentaires pour chacun des gènes et des familles de protéines. L'amélioration apportée par MAGUS est la possibilité d'annoter, à partir d'une seule page de résultats, les différents modèles de gènes regroupés en familles de protéines calculées d'après une analyse *in silico*.

La page d'un locus (une région d'ADN chromosomique ayant un ou plusieurs modèles de gènes chevauchants) visualise, grâce au GBrowse, l'ensemble des modèles de gènes, ainsi que le gène validé s'il existe. Si ce n'est le cas, l'annotateur peut se référer à la liste des modèles ordonnés selon la taille de la protéine prédite à partir de chaque modèle et le pourcentage

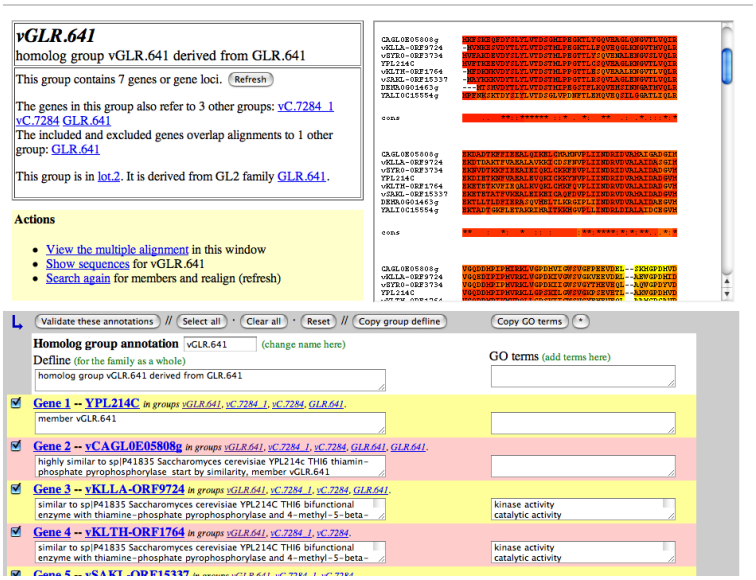
[1] MAGUS : http ://magus.gforge.inria.fr/

de similarité avec *S. cerevisiae*. En sélectionnnant un modèle, l'annotateur accède à la page d'annotation pour ce modèle (cf. fig. 5.2). Cette page se compose de 4 parties. Dans la partie supérieure de la page, se trouve la plupart des renseignements utiles pour l'annotateur et leurs accès, répartis dans trois cadres. Comme nous le voyons sur la figure 5.2, le cadre supérieur droit, également présent sur la page du locus, indique l'ensemble des modèles de gènes possibles (en rouge) pour ce locus, ainsi que l'analyse GeneMark, les alignements de type BLASTx de ces modèles contre diverses banques de données (dont les génomes en cours d'annotation) et le voisinage de ce locus. Le modèle de gène surligné en jaune est le modèle regardé par l'annotateur. Le cadre supérieur gauche décrit le gène prédit choisi par le système (puis par l'annotateur), ainsi que quelques analyses pour ce modèle et son produit de traduction. Dans le cadre inférieur gauche (fond jaune), l'annotation du modèle de gène choisi par le système est indiqué dans une zone de texte libre. Des liens internes sur la page ('Quick links') ou externes ('Results') donnent accès aux analyses *in silico* disponibles. La partie inférieure de la page d'annotation (non montrée ici) regroupe les zones de saisie de texte libre pour l'annotation et un commentaire facultatif, ainsi que les séquences (ADN, ARNm et protéine) pour le modèle de gène choisi, avec la possibilité de faire, pour ladite séquence, une analyse BLAST contre les banques de données proposées par Génolevures.

Comme la version précédente du système d'annotation basé sur CAAT-Box, en permettant l'annotation dans un contexte de génomique comparée, MAGUS assure une homogénéité et une cohérence de l'annotation, tout en facilitant le travail des annotateurs. Mais la première particularité de MAGUS est l'annotation de loci homologues issus de plusieurs génomes à partir d'une seule page (cf. fig. 5.3). Les résultats d'analyses *in silico* telles que les famille de protéines, alignements multiples, permettent d'identifier quels modèles de gène (un seul modèle par locus) sont susceptibles d'être homologues. Ces modèles forment alors un groupe d'homologues. Cette annotation de groupes d'homologues présente divers avantages. Tout d'abord, le gain de temps est certain car l'annotateur voit rapidement quels modèles de gènes peuvent être validés et avec quelle annotation. Ensuite, l'annotation est plus homogène car les membres d'un groupe partagent assurément la même annotation, ou au moins le même genre d'annotation. En effet, l'annotateur peut vérifier, sur une seule et même page, que l'annotation fonctionnelle est cohérente pour tous les membres du groupe. Si ce n'est pas le cas, il peut regarder plus en détail la composition du groupe, enlever éventuellement le ou les membres qu'il juge non homologue(s), décomposer le groupe en sous-groupes. De plus, l'annotatuer visualise la structure des différents membres du groupe, à savoir présence ou non d'intron pour ces membres. Cela permet de préférer un modèle de gène plutôt qu'un autre.

La seconde particularité de MAGUS est l'observation de la conservation de la synténie (*i.e.* l'ordre des gènes sur une région d'ADN) pour les différents homologues d'un groupe. Comme nous le voyons sur la figure 5.4 représentant la partie basse de la page d'annotation de gènes homologues, les voisinages de chacun des membres (surlignés en jaune car ce sont les modèles de gène choisis par l'annotateur) du groupe, sont disposés verticalement, chaque voisinage centré sur le modèle de gène choisi. L'étude de la synténie peut ainsi confirmer ou non l'appartenance d'un locus (et donc d'un gène) à un groupe d'homologues. Cela permet aussi de visualiser éventuellement l'absence d'un gène dans le voisinage, par comparaison avec

FIG. 5.3 – Partie supérieure de page d'annotation d'un groupe de modèles de gène homologues par le système MAGUS.



ce qui a été validé sur les voisinages des autres modèles homologues.

L'annotation des levures du projet Génolevures 3, tout comme celles des levures de Génolevures 2, est ainsi rapide et de bonne qualité. Cependant, pour une qualité accrue, il faut une vision d'ensemble sur le génome annoté en tant que système fonctionnel d'un point de vue biologique. L'application des règles définies au chapitre précédent nous permet d'atteindre ce niveau accru de qualité. Certaines règles sont d'ailleurs entièrement intégrées dans le système MAGUS.

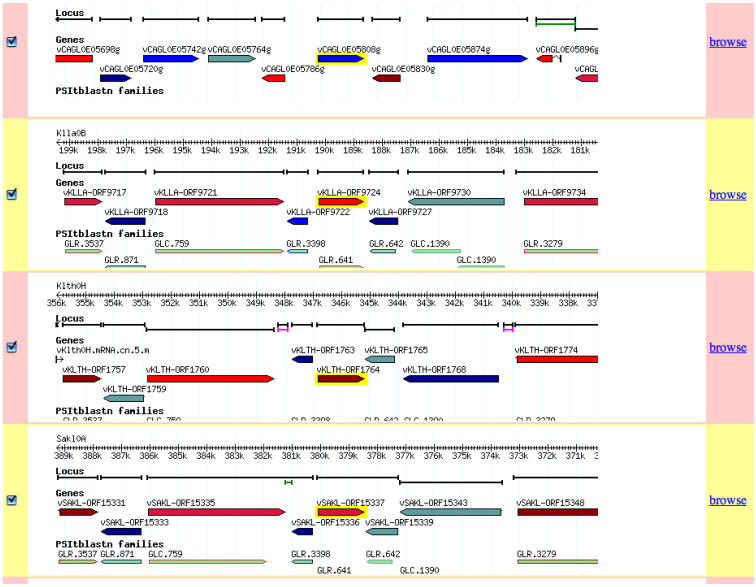
5.2 Application des règles de cohérence

Nous présentons dans un premier temps les sources de données utilisées pour l'application de certaines des règles de cohérence (*e.g.* la règle 4, p. 85, validant les motifs aux bornes introniques). Nous présentons ensuite la mise en œuvre de certaines des règles de cohérence définies dans le chapitre 4 ainsi que les résultats de leur applications.

5.2.1 Données pour l'élaboration des règles

J'ai défini ces règles à partir de la littérature scientifique, d'analyses de séquences, d'études bio-informatiques. Lorsque les informations sont collectées à partir de bases de données ou d'analyses bio-informatique, un problème majeur doit être considéré : celui de leur validité. En effet, les bases de données utilisent les données d'autres bases et la mise à jour d'une base n'entraîne pas obligatoirement celle des autres. Il faut donc recouper ces informations, tenir compte des changements effectués (par exemple, la disparition d'un gène prédit) afin d'obtenir

FIG. 5.4 – Visualisation de la conservation de la synténie pour les membres d'un groupe d'homologues (partie inférieure de page d'annotation d'une groupe de gènes homologues.



un jeu de données le plus juste à un moment donné.

Pour un gène annoté dans un nouveau génome, nous disposons de son annotation Génolevures et de son appartenance éventuelle à une famille de protéines. Pour les données concernant *S. cerevisiae*, j'ai fait appel à plusieurs sources de données : familles de protéines, gènes essentiels, interactions protéiques, connaissances biologiques.

5.2.1.1 Motifs introniques

Les motifs introniques sont déterminés par l'étude thématique Génosplicing, coordonnée par Cécile Neuvéglise [Neuvéglise, 2005]. Une recherche systématique d'intron potentiel est réalisée pour chacune des nouvelles espèces à partir des motifs connus de *S. cerevisiae* et des autres levures. L'étude Génosplicing, est accessible sur le site internet Génolevures [2].

5.2.1.2 Connaissances bibliographiques

Les données de séquences relatives à une espèce donnée sont prises sur la base de données UniProt [Apweiler et al., 2007]. Cette base de données généraliste est en effet régulièrement mise à jour et intègre toute information relative à une séquence nucléique. L'interrogation de la base de données est réalisée grâce au logiciel SRS ("Sequence Retrieval System") [Etzold et al., 1996], système de navigation et d'interrogation de diverses banques de données utilisées en biologie moléculaire.

[2] http ://cbi.labri.fr/Genolevures/genosplicing/

5.2.1.3 Famille de protéines

Lors de la phase Génolevures 2 (annotation de quatre nouveaux génomes et comparaison avec *S. cerevisiae*), les protéines des cinq levures ont été analysées de façon à être regroupées en familles de protéines homologues[3] [Nikolski and Sherman, 2007]. Trois types de familles sont définis :

- les familles standards ou familles GLS ("Génolevures Standard"),
- les familles robustes ou familles GLR ("Génolevures Robust") et
- les familles consensus ou familles GLC ("Génolevures Consensus").

Les familles GLS ont été définies manuellement à partir de recherches bibliographiques. Elles concernent des protéines assurant des fonctions vitales pour la cellule. Les familles GLR et GLC ont été calculées selon un algorithme de clustering consensus [Nikolski and Sherman, 2007]. Ces familles sont maintenues à jour au fil de l'avancement de l'annotation de nouveaux génomes. En effet, leur composition étant basée sur des alignements de séquences, elles peuvent être calculées dès la phase d'annotation syntaxique (détermination des CDS).

Une famille Génolevures se définit par deux caractéristiques :

- son profil phylétique : il correspond à la présence ou l'absence d'au moins une protéine dans chacune des espèces considérées, celles-ci étant ordonnées selon l'arbre phylogénique présenté § 3.3 fig. 3.6 : par exemple "sckdy" signifie qu'il y a au moins une protéine dans les espèces *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii*et *Y. lipolytica*.
- son profil phylogénique : il correspond au nombre de protéines pour chacune des espèces considérées, ordonnées comme précédemment.

Les familles Génolevures me permettent d'inférer des propriétés sur les nouveaux génomes à partir de celles de *S. cerevisiae* (quand elles sont présentes dans une famille). J'ai utilisé les familles calculées pour Génolevures 2 (5 espèces, version du 31/01/2006 [4]).

5.2.1.4 Fonctions essentielles

Diverses études expérimentales ont été réalisées chez *S. cerevisiae* afin de déterminer les gènes essentiels. Les données pertinentes sont ensuite regroupées puis triées et recoupées afin d'obtenir un ensemble de gènes essentiels le plus complet et le plus juste possible. Néanmoins, il ne faut pas oublier que ces gènes sont dits "essentiels" pour *S. cerevisiae*, ce qui n'implique pas qu'ils le sont forcément pour les autres espèces de levures. Par ailleurs, chaque espèce a pu développer un ou plusieurs gènes essentiels uniquement pour elle.

La première source de gènes essentiels provient de la base SGD. Les informations présentes sur cette base de données référence pour *S. cerevisiae*, sont maintenues à jour et validées grâce à un travail de curation manuelle. Des expériences biologiques de délétion systématique de gène (un gène cible connu est supprimé) ont été réalisées chez *S. cerevisiae*. L'observation d'un phénotype non viable de la cellule permet de qualifier le gène supprimé comme létal. SGD a ainsi répertorié 1100 gènes létaux donc potentiellement essentiels.

[3] Étude des familles de protéines pour le projet Génolevures : http ://cbi.labri.fr/Genolevures/fam/index.html.
[4] http ://cbi.labri.fr/Genolevures/download/GL2_PF.php

Un deuxième ensemble de gènes essentiels provient de la base de données CYGD du MIPS [5]. Le résultat d'expériences de délétion isolée ou à grande échelle a, là aussi, permis d'identifier 949 gènes létaux, et 160 gènes "létaux/viables" (leur délétion, dans au moins deux souches de levure, a des résultats différents selon la souche).

Le projet de délétion systématique du génome de *S. cerevisiae*, "*Saccharomyces* Genome Deletion Project" [Giaever et al., 2002], regroupant des centres universitaires nord-américains et européens, cherche à définir la fonction d'ORF en étudiant les conséquences phénotypiques de la délétion d'un gène cible. En mai 2007, les expériences réalisées par ce projet ont permis la caractérisation de 1 122 gènes essentiels.

La base de données DEG [6] [Zhang et al., 2004] présente les gènes essentiels pour plusieurs organismes dont *S. cerevisiae*. Leurs données sur la levure proviennent de la base CYGD du MIPS mais elles n'étaient pas à jour au 23/08/2007. En effet, DEG ne contient ni l'ensemble des gènes du MIPS (seuls 870 gènes sont présents) ni les autres grands projets d'identification de gènes essentiels tels que ceux cités ci-dessus. Pour ces raisons, j'ai jugé ces données trop peu pertinentes et n'en avons pas tenu compte pour la constitution de notre ensemble de gènes essentiels.

Finalement, j'ai constitué un ensemble de 1 124 gènes essentiels.

5.2.1.5 Interactions et complexes protéiques

La base de données IntAct [Hermjakob et al., 2004b] regroupe les interactions protéiques pour *S. cerevisiae* mise en évidence au cours de différents projets d'analyse à grande échelle [Uetz et al., 2000, Ito et al., 2001, Ho et al., 2002]. La banque de données BioGrid[7] [Stark et al., 2006] groupe Elle aussi des données d'interactions protéine-protéine issues de travaux complémentaires [Aalto et al., 1993, Krogan et al., 2004].

J'ai choisi aussi de prendre trois complexes protéiques qui sont l'oxydase du cytochrome C (9 protéines), l'ATP synthase F1F0 (8 protéines) et le protéasome 26S (31 protéines). J'ai choisi ces complexes car ils nous offrent également un support pour la prédiction des interactions protéine-protéine et je reprendrai leur analyse dans le chapitre 6 sur la mise en évidence des IPP par méthode expérimentale. Ces trois complexes sont soumis à une très forte pression sélective car indispensable à la cellule. En conséquence, ils sont très conservés au sein des espèces d'une même branche phylogénique telle que celles des levures.

5.2.2 Langage et environnement

Nous avons choisi de programmer l'ensemble des règles de cohérence dans le langage Perl ("Practical Extraction and Report Language"), créé par Larry Wall en 1987. Le choix de ce langage s'est imposé car nous utilisons la bibliothèque BioPerl qui permet la manipulation de nos données par un modèle objet.

[5] Comprehensive Yeast Genome Database, la base de données consacrée à *S. cerevisiae*, proposée par le Munich Information Center for Protein Sequence : http ://mips.gsf.de/genre/proj/yeast/

[6] Database of Essential Genes : http ://tubic.tju.edu.cn/deg/

[7] BioGRID : "General Repository for Interaction Datasets" : http ://www.thebiogrid.org/.

Certains formats, tels que les formats FASTA et EMBL, ont su s'imposer et sont largement utilisés par les domaines de la biologie et de la bio-informatique, permettant un échange de données rapide. Le format FASTA, créé pour le logiciel FASTA [Pearson and Lipman, 1988] est ainsi repris comme format d'entrée des données pour les outils d'alignement de séquences simple (BLAST [Altschul et al., 1990]) ou multiple (CLUSTAL W [Higgins et al., 1994], T-Coffee [Notredame et al., 2000]).

Ainsi, des outils codés en Perl et dédiés aux domaines de la bioinformatique, de la génomique et des Sciences de la Vie, sont développés dans le cadre du projet BioPerl [8], association internationale ouverte de développeurs officielle depuis 1995. BioPerl fournit de très nombreux modules [9] permettant la manipulation d'objets biologiques (séquence, alignement, base de données), dans les formats communément utilisés (par exemple les formats BLAST, FASTA, EMBL, OBO [10]), l'exécution et le traitement informatique de résultats de logiciels bio-informatiques (par exemple BLAST, T-Coffee).

La base de données suit le schéma BioSQL [11], une base de données orientée objet permettant de gérer des séquences biologiques et leurs attributs et relations. Le module bioperl_db réalise une couche d'adaptation objet-relationnel : les objets BioPerl sont stockés dans la base de données et la couche se charge de la conversion entre les deux représentations.

Le système d'annotation et la visualisation des données par le GBrowse utilisent le module principal `Bio : :DB : :Seqfeature` et ses modules dérivés en raison de la meilleure prise en charge de ce dernier. Cette base de données peut être extraite de la base de données BioSQL.

Les règles de cohérence sont appliquées aux données de Génolevures soit directement sur la base de données soit sur des fichiers plats issus de la base de données. La base de données étant modifiée par les annotateurs pendant l'application des règles, il s'avère plus facile d'appliquer celles-ci sur un fichier de données issues de la base à un instant `t`.

Les incohérences révélées par les règles sont stockées sous forme de fichiers plats. Les règles ne modifient pas directement les données dans la base. En effet, l'annotation étant manuelle et/ou semi-automatique, les incohérences doivent être soumises à l'appréciation des annotateurs avant toute modification dans la base de données. Seules les corrections apportées au niveau de la syntaxe des annotations, peuvent être directement intégrées à la base de données.

5.2.3 Règles élémentaires

Pour rappel, les règles (R) de cohérence appliquées au niveau de l'élément génique sont les suivantes :

- R1 : nombre d'exon et d'intron,
- R2 : alternance exon / intron,
- R3 : contiguïté des bornes exoniques et introniques,

[8] http ://www.bioperl.org/wiki/Main_Page

[9] http ://doc.bioperl.org/releases/bioperl-1.5.2/

[10] Le site Open Biological Ontology répertorie les ontologies disponibles dans le domaine de la biologie ; il a développé : format de données utilisé par la GO

[11] http ://www.bioperl.org/wiki/BioSQL

- R4 : motifs introniques,
- R5 : codon initiateur et codon terminateur,
- R6 : cadre de lecture,
- R7 : longueur minimale de séquence codante et
- les règles syntaxiques.

5.2.3.1 Architecture du gène

L'implémentation de R1 à R7 suit la définition formelle pour chacune de ces règles et ne pose pas de problème algorithmique. Par conséquent, elle n'est pas présentée dans ce document.

La règle 4 n'est pas appliquée pour les motifs des points de branchement car ce motif n'est pas conservé dans la base de données actuellement.

Les règles R5, R6 et R7 sont vérifiées en même temps mais indépendamment les unes des autres de façon à savoir quelle règle n'est pas satisfaite par le gène codant la protéine.

Résultats Au 20/10/07, les erreurs détectées concernant le respect des contraintes pour les gènes à intron (R1, R2, R3 et R4) ont en fait été corrigées au fur et à mesure de l'annotation des génomes. De même, les motifs introniques non valides sont re-soumis aux annotateurs afin d'être confirmés ou non. Si le motif s'avère faux, la prédiction de l'intron est enlevée et le locus (région d'ADN correspondant à un ORF ou plusieurs se chevauchant et dans laquelle l'intron était prédit) est à réannoter. Si le motif intronique est jugé correct par les annotateurs, il enrichit la liste des motifs déjà valides. Après correction à partir des analyses R3– R4, il reste un gène de *C. glabrata* sans codon initiateur (CAGL0K03459g).

5.2.3.2 Syntaxe de l'annotation

La vérification de la syntaxe d'annotation se base sur les règles établies par le consortium Génolevures pour la rédaction de l'annotation d'un modèle de gène validé codant une protéine. La présence d'une annotation est vérifiée pour chaque gène validé. La vérification orthographique s'effectue sur la description et le niveau de similarité de la séquence. Des expressions régulières reconnaissent la structure du niveau de similarité et la corrigent si besoin. L'existence du gène de *S. cerevisiae* est également vérifiée vis-à-vis de ceux présent dans la banque de données SGD.

Nous présentons ici quelques exemples de corrections réalisées, codée en Perl, illustrées par des cas réels.

Correction orthographique des niveaux de similarités Chaque CDS potentiel traduit en protéine et ayant une similarité, d'après alignement de type BLASTp avec une protéine connue, se voie attribuer un niveau de similarité selon la qualité de l'alignement. Ce niveau de similarité correspond à un élément possible du vocabulaire contrôlé défini ci-après : "`some similarities with`", "`weakly similar to`", "`similar to`" et "`highly similar to`".

L'extrait de code présenté au listing 5.1 montre comment s'effectue la vérification orthographique pour les niveaux de similarité "`highly similar to`" et "`similar to`". La vérification

est faite pour tout couple formé d'un CDS et de son annotation non nulle (ligne 7) (annotation non nulle vérifiée lors d'une étape précédente non montrée ici).
Les lignes 9 à 14 concernent la vérification pour "`highly similar to`". Si l'annotation commence par des mots ressemblant à "`highly similar`" (ligne 9) mais n'est pas correcte pour ce niveau de similarité (ligne 10), les deux premiers mots sont correctement réécrits (ligne 11). Si l'annotateur a écrit "`with`" au lieu de "`to`", le remplacement est effectué (ligne 12). Puis la préposition "`to`" est rajoutée si l'annotateur l'a ommise (ligne 13). Nous obtenons ainsi :

- *avant correction :* le CDS `YALI0F20768g` a pour annotation "`highly similar sp|P32495 Saccharomyces cerevisiae YDL208w NHP2 nucleolar rRNA processing protein`",
- *après correction :* le CDS `YALI0F20768g` a pour annotation "`highly similar to sp|P32495 Saccharomyces cerevisiae YDL208w NHP2 nucleolar rRNA processing protein`".

Les lignes 16 à 19 concernent la vérification pour "`similar to`". Si l'annotation commence par des mots ressemblant à "`similar`" mais n'est pas correcte pour ce niveau de similarité (ligne 16), le premier mot est correctement réécrit (ligne 17). Puis la préposition "`to`" est rajoutée si l'annotateur l'a ommise (ligne 18).

Listing 5.1 – Orthographe du niveau de similarité dans l'annotation d'un CDS.

```
1  # table de hachage pour l'espece consideree, composee des couples
2  #{clef = nom du CDS, valeur = annotation}
3
4  my %CDS;
5  # remplissage de la table %CDS
6
7  while (($cds, $annotation) = each %CDS) {
8      # niveau de similarite: 'highly similar to'
9      if (($annotation =~ /^[highly]* \s [simlar]* /ix)
10         && ($annotation !~ /^highly similar to/)) {
11         $annotation =~ s/^\w+\s\w+\s/highly similar /;
12         $annotation =~ s/^(\w+\s\w+)\s(with)\s*/$1 to /;
13         $annotation =~ s/^(\w+\s\w+)\s(.[^to]+)/$1 to $2/;
14     }
15     # niveau de similarite: 'similar to'
16     elsif (($annotation =~ /^[simlar]* \s /ix) && ($annotation !~ /^similar to/)) {
17         $annotation =~ s/^\w+\s/similar /;
18         $annotation =~ s/^(\w+)\s(.[^to]+)/$1 to $2/;
19     }
20 }
```

Nous obtenons ainsi :

- *avant correction :* le CDS `DEHA2E06072g` a pour annotation "`similar CA3407|IPF9406 Candida albicans IPF9406 unknown function`",
- *après correction :* le CDS `DEHA2E06072g` a pour annotation "`similar to CA3407|IPF9406 Candida albicans IPF9406 unknown function`".

Aux lignes 11 à 13, 17 et 18, l'opérateur de substitution **"s//"** est appliqué à l'annotation `$annotation`. Les séparateurs **"/"** délimitent :

- en première position, une expression régulière à laquelle peut correspondre une partie de la chaîne de caractères (ici l'annotation),

– en seconde position, une chaîne de caractères qui remplacera la sous-chaîne de caractères qui correspondait à l'expression régulière précédante.

La vérification orthographique des autres niveaux de similarité et de certains noms d'espèces suit le même principe.

Structure de l'annotation lors d'une similarité avec *S. cerevisiae* Lorsqu'un CDS est prédit similaire à un gène de *S. cerevisiae*, l'annotation doit avoir la structure suivante : "⟨*niveau de similarite*⟩ ⟨*identifiant de S. cerevisiae*⟩ ⟨*Saccharomyces cerevisiae*⟩ ... \... ...\...⟨*nom systematique de S. cerevisiae*⟩ ⟨*suite de l'annotation*⟩".

Soit, par exemple :
"highly similar to sp|P32495 Saccharomyces cerevisiae YDL208w [...]".

Nous présentons ici deux étapes de vérification du respect de cette structure :

– présence du nom systématique *après* le nom de l'espèce *"Saccharomyces cereviae"* et
– présence de l'identifiant UniProt *avant* le nom de l'espèce *"Saccharomyces cereviae"*.

Présence du nom systématique après le nom de l'espèce L'extrait de code du listing 5.2 présente la gestion de la place du nom systématique du gène de *S. cerevisiae*.

Listing 5.2 – Positionnement du nom systématique après le nom de l'espèce.

```
1  # table de hachage pour l'espece consideree, composee des couples
2  # {clef = nom du CDS, valeur = annotation}
3  my %CDS;
4  # remplissage de la table %CDS
5
6  while (($cds, $annotation) = each %CDS) {
7      # place du nom systematique de S.cerevisiae apres son nom d'espece
8      if ($annotation =~
9          /^((highly similar to)|(similar to)|(weakly similar to)|
10             (some similarities with)) Y[A-P](L|R)\d+[cwCW]/) {
11         (my $NomSyst) =  ($annotation =~ / (Y[A-P][LR]\d+[cwCW]+\-?\w?) \s [\w\|]+/x);
12         if ($NomSyst) {
13             $annotation =~ s/$NomSyst//;
14             if ($annotation !~ $NomSyst ) {
15                 $annotation =~
16                     s/Saccharomyces cerevisiae /Saccharomyces cerevisiae $NomSyst /;
17             }
18         }
19     }
20 }
```

Pour chaque CDS ayant une annotation (ligne 6), si l'annotation contient le nom systématique placé après le niveau de similarité (lignes 8-10), ce nom, correspondant à une chaîne de caractères donnée (ligne 11), est supprimé lors de sa première occurrence (ligne 13). Si ce nom n'est plus présent dans l'annotation (ligne 14), il est rajouté après le nom d'espèce (lignes 15-16).

L'extrait de code du listing 5.3 présente la gestion de la présence du nom systématique du gène de *S. cerevisiae*. Ce test intervient après celui présenté listing 5.2.

Pour chaque CDS ayant une annotation (ligne 11), ce test s'applique pour une annotation contenant un niveau de similarité avec *S. cerevisiae* (lignes 13-15) sans nom systématique de gène après le nom d'espèce (ligne 16). L'identifiant UniProt présent dans l'annotation (ligne 17) est alors utilisé pour retrouver le nom du gène (ligne 18) et l'ajouter ainsi après le nom d'espèce (lignes 19-20). Si l'identifiant UniProt n'est pas connu dans les données initiales (table %SystId, ligne 8), un message d'erreur avertit l'utilisateur (lignes 22-23).

Listing 5.3 – Présence du nom systématique après le nom de l'espèce.

```
1  # table de hachage pour l'espece consideree, composee des couples
2  # {clef = nom du CDS, valeur = annotation}
3  my %CDS;
4  # remplissage de la table %CDS
5
6  # table de hachage pour S. cerevisiae, composee des couples {clef =
7  # nom systematique, valeur = Identifiant UniProt}
8  my %SystId;
9  # remplissage de la table %SystId
10
11 while (($cds, $annotation) = each %CDS) {
12     # presence du nom systematique de S. cerevisiae
13     if ($annotation =~ /^((highly similar to)|(similar to)|(weakly similar to)|
14                           (some similarities with))
15                           (\w+\|+\w+) (Saccharomyces cerevisiae)/
16         && $annotation !~ /Saccharomyces cerevisiae Y[A-P](L|R)\d+[cwCW]/) {
17         (my $Ident) = ($annotation =~ /([\w\|]+) Saccharomyces cerevisiae/);
18         if ($Ident && $SystID{$Ident}) {
19             $annotation =~
20                 s/Saccharomyces cerevisiae/Saccharomyces cerevisae $SystID{$Ident}/;
21         } else {
22             print STDERR "PBL: pas de nom systematique pour $Ident: ",
23             $cds, " ", $annotation,"\n";
24         }
25     }
26 }
```

Présence de l'identifiant UniProt avant le nom de l'espèce L'extrait de code du listing 5.4 présente la gestion de la présence et la place de l'identifiant UniProt pour le gène de *S. cerevisiae*.

Pour chaque CDS ayant une annotation (ligne 11), ce test s'applique pour une annotation contenant un niveau de similarité avec *S. cerevisiae* (lignes 13-14) sans identifiant UniProt. Le nom systématique présent dans l'annotation est alors utilisé (lignes 15-16) pour retrouver l'identifiant correspondant (ligne 17) et l'ajouter ainsi avant le nom d'espèce (ligne 18). Si l'identifiant UniProt n'est pas connu dans les données initiales (table %SystId, ligne 8), un message d'erreur averti l'utilisateur (lignes 20-21).

Par exemple, nous obtenons ainsi :

- *avant correction :* le CDS `YALI0D26191g` a pour annotation "`weakly similar to Saccharomyces cerevisiae YLR337c VRP1 verprolin`",
- *après correction :* le CDS `YALI0D26191g` a pour annotation "`weakly similar to sp|P37370 Saccharomyces cerevisiae YLR337c VRP1 verprolin`".

Listing 5.4 – Présence de l'identifiant UniProt avant le nom d'espèce.

```
# table de hachage pour l'espece consideree, composee des couples
# {clef = nom du CDS, valeur = annotation}
my %CDS;
# remplissage de la table %CDS

# table de hachage pour S. cerevisiae, composee des couples {clef =
# Identifiant UniProt, valeur = nom systematique}
my %UniProtSyst;
# remplissage de la table %UniProtSyst

while (($cds, $annotation) = each %CDS) {
    # verifier la presence de l'identifiant UniProt a partir du nom systematique
    if ($annotation =~ /^((highly similar to)|(similar to)|(weakly similar to)|
            (some similarities with)) (Saccharomyces cerevisiae Y[A-P](L|R)\d+[cwCW]/) {
        (my $NomSyst) = ($annotation =~ /Saccharomyces \s cerevisiae \s
                        (Y[A-P](L|R)\d+[cwCW]+\-?\w*) \s/x);
        if ($UniProtSyst{$NomSyst}) {
            $annotation =~ s/(Saccharomyces cerevisiae)/$UniPortSyst{$NomSyst} $1/;
        } else {
            print STDERR "PBL: pas d'identifiant UniProt pour $NomSyst ",
                $cds, "\t", $annotation,"\n";
        }
    }
}
```

L'ensemble des vérifications est itéré trois fois sur la même annotation. En effet, les tests sont ordonnés selon la progression de la lecture de l'annotation. Mais lors de la première itération, la correction d'une erreur par un test peut révéler une seconde erreur qui peut se corriger par un test déjà passé. Cette seconde erreur sera alors corrigée lors de la deuxième itération. Pour les fichiers d'annotations soumises, aucune correction automatisable n'était effectuée lors de la troisième itération (cf. tab. 5.1). Cependant, par précaution, j'ai gardé cette dernière itération.

Résultats La vérification de la syntaxe de l'annotation a été appliquée sur les quatre espèces de Génolevures 2. Lors d'une première phase d'annotation de Génolevures 2, la vérification de la syntaxe avait permis de corriger environ 2/3 des annotations soit environ 16 000 gènes. Le tableau 5.1 présente le nombre de corrections détectées pour les différents fichiers d'annotations lors de chacune des itérations des tests de vérification, ainsi que le nombre d'annotations ayant une ou plusieurs erreurs persistantes.

Tab. 5.1 – Vérification syntaxique des annotations Caractéristiques des levures Génolevures 2. annotations = nombre d'annotations ; IX = pour l'itération X, nombre d'erreurs détectées ; err. = erreurs persistantes (données Génolevures, 01/01/2008).

espèce	annotations	I1	I2	I3	err.
C. glabrata	5 204	5 064	26	26	3
K. lactis	5 084	8 439	1 299	1 260	666
D. hansenii	6 277	6 191	181	174	124
Y. lipolytica	6 424	6 911	597	590	92

Ces erreurs sont étudiées manuellement et éventuellement corrigées soit directement soit après soumission aux annotateurs.

5.2.4 Règles chromosomiques

Les règles appliquées sont les règles vérifiant l'absence de chevauchement entre éléments génomiques et la présence d'un seul centromère par chromosome.

L'implémentation de ces deux règles suit leur définition donnée dans le chapitre précédent. Ne présentant pas d'intérêt algorithmique, elle ne sera pas présentée dans ce document.

5.2.4.1 Absence de chevauchement

Les cas de chevauchement détectés ont été signalés aux annotateurs au fur et à mesure de l'avancement du travail d'annotation. A l'heure actuelle, il ne subsiste plus de chevauchement pour les espèces étudiées lors du projet Génolevures 2.

5.2.4.2 Unicité du centromère

La présence d'un seul centromère par chromosome a été vérifiée pour chaque espèce. Les chromosomes de *C. glabrata*, *K. lactis* et *Y. lipolytica* comptent bien chacun un centromère. Seule *D. hansenii* n'a aucun centromère identifié. La raison est que les séquences consensuelles caractéristiques du centromère ne sont pas présentent chez cette levure. À notre connaissance, les biologistes n'ont pas identifié les centromères chez *D. hansenii*.

5.2.5 Règles génomiques

Les règles génomiques concernent la présence des fonctions essentielles, la composition complète de complexes protéiques par homologie avec ceux observés chez *S. cerevisiae*, ainsi que l'intégration des connaissances biologiques.

5.2.5.1 Fonctions essentielles

La validation de la présence de fonctions essentielles concerne ainsi la détection des gènes létaux et synthétiques létaux, ainsi que le détection des gènes impliqués dans des complexes vitaux (cf. § 5.2.1.4). L'approche utilisée est la même pour ces deux catégories de gènes.

La présence des gènes essentiels de *S. cerevisiae*, ensemble constitué à partir des diverses source de données, est vérifiée par alignement de séquences, avec le programme BLASTp, dans l'ensemble des gènes prédits lors de l'annotation du génome.

Pour une espèce considérée, les séquences protéiques prédites à partir des modèles de gènes validés sont extraites de la base de données SeqFeature au format FASTA. Elles constituent la banque de données dans laquelle les séquences protéiques essentielles de *S. cerevisiae* sont recherchées.

L'alignement entre une séquence protéique de *S. cerevisiae* et celle de l'espèce en cours, est jugé valide si deux conditions sont satisfaites : critères :

1. la longueur de l'alignement est supérieure ou égale au seuil fixé,

2. la e-valeur (probabilité que cette homologie ne soit pas due au hasard) est inférieure ou égale à la e-valeur seuil fixée.

Ces seuils de longueur d'alignement et de e-valeur doivent être adaptés en fonction de l'espèce considérée, afin de tenir compte de la distance évolutive entre cette espèce et l'organisme de référence, ici *S. cerevisiae*.

Par exemple, pour *Y. lipolytica*, j'ai fixé le seuil de longueur d'alignement à 50% et la e-valeur à 1.0×10^{-7}. Ces valeurs ont été choisies par apprentissage sur le jeu test composé des gènes de la cytochrome C oxydase. En effet, des analyses préliminaires montraient la présence de ces gènes. Les valeurs de seuillage choisies doivent nous permettre de détecter ces gènes.

Les gènes essentiels de *S. cerevisiae* non présents dans l'annotation doivent alors être recherchés sur la séquence génomique ADN de l'espèce considérée par tBLASTn.

Résultats La présence des 1 124 gènes essentiels chez *S. cerevisiae* a été recherchée pour *Y. lipolytica*. En utilisant comme seuils 50% d'homologie pour l'alignement et une e-valeur maximale de 1.0×10^{-7}, 1 011 gènes ont une séquence homologue à un gène prédit de *Y. lipolytica*. Parmi les 49 gènes non retenus selon ces critères, 41 ont un alignement avec une e-valeur supérieure à 0.1, ce qui traduit un alignement peu fiable.

5.2.5.2 Complexes protéiques

Les complexes protéiques (cytochrome C oxydase, ATP synthase F1F0 et protéasome) étudiés ici sont d'un intérêt vital pour la cellule. Aussi les gènes codant leurs protéines sont inclus dans l'ensemble des gènes essentiels. Mais cela n'est pas redondant. D'une part, la recherche des gènes essentiels permet d'avoir une analyse à grande échelle de la qualité de l'annotation réalisée. D'autre part, regarder plus précisément un complexe protéique permet de voir ce qu'il manque d'un point de vue fonctionnel. Il s'avère qu'ici, nous avons pris des complexes essentiels mais l'étude de complexes non essentiels permettrait de juger de la spécificité biologique de l'organisme considéré par rapport à l'organisme de référence.

Résultats Les résultats concernant la recherche des complexes protéiques choisis pour les quatre espèces sont présentés dans la table 5.2. Seule l'annotation de *Y. lipolytica* ne contient pas un gène du cytochrome C oxydase et un autre de l'ATP synthase F1F0.

J'ai voulu savoir si l'absence des gènes des complexes était réelle, due à l'évolution de *Y. lipolytica* par rapport aux autres levures, ou simplement un oubli dans l'annotation.

À la recherche de gènes manquants... Chez *Y. lipolytica*, le gène manquant pour la cytochrome C oxydase serait le gène homologue au gène COX9/YDL067C (gène sans intron) qui code la protéine Cox9p d'une longueur de 59 aa. La recherche par tBLASTn (matrice BLOSUM62) de la séquence protéique de Cox9p sur le génome de *Y. lipolytica* m'a permis d'identifier ce gène, baptisé YALI0F04114g. Ce gène, composé de deux introns, code une protéine de 61 aa (vérification expérimentale par ADNc par C. Neuvéglise, résultats non publiés). La composition de ce gène est plutôt atypique à plusieurs points de vue :

– le gène a une longueur de 895 nt mais la séquence codante est seulement de 183 nt,

Tab. 5.2 – Conservation des complexes protéiques du cytochrome C, de l'ATP synthase et du protéasome, essentiels pour *S. cerevisiae* chez les autres levures du projet Génolevures 2.

	cyt. C oxy (9 g.)	ATP synthase (8 g.)	protéasome (31 g.)
CAGL	9	8	31
KLLA	9	8	31
DEHA	9	8	31
YALI	8	7	31

– l'exon 2 a une taille de 6 nt.

Ce gène n'a pas été annoté comme prédiction valide lors de la phase d'annotation de Génolevures 2. Pourtant le plus grand exon (110 nt) satisfait la contrainte de taille minimale fixée à 80 nt. Peut-être que l'annotateur a jugé cette séquence trop courte et sans codon start comme faux positif...

J'ai ensuite recherché le gène homologue au gène ATP15/YPL271W codant la sous-unité ϵ de l'ATP synthase F1F0 d'une longueur de 62 aa. Une première recherche par alignement de type tBLASTn me permet de le localiser sur le chromosome A de *Y. lipolytica*. L'alignement réalisé ne porte que sur la seconde partie de la séquence protéique. Nous suspectons, là aussi, la présence d'au moins un intron. La recherche doit être affinée, notamment en utilisant une matrice de distance pour l'alignement de type tBLASTn plus appropriée telle que la matrice PAM50. La présence de ce gène nous semble d'autant plus réelle que le gène YPL271w appartient à la famille GLC.2073, de profil phylétique "sckd-" et de profil phylogénétique "1 1 1 1 0".

5.2.5.3 Intégration des données bibliographiques

Les gènes connus issus de UniProt sont recherchés parmi les gènes prédits par l'annotation : la recherche se fait sur l'identifiant du gène. Les gènes connus mais non identifiés en tant que tels dans l'annotation sont recherchés par alignement de type BLASTp avec les modèles de gènes. L'alignement est valide s'il satisfait les contraintes suivantes : 90% de similarité de séquence pour la longueur de l'alignement et une e-valeur maximale de 1.E-40. Les gènes qui ne sont pas localisés doivent alors être recherchés sur la séquence génomique de l'espèce par alignement de type tBLASTn. Lors des recherches d'homologie de séquences, nous vérifions qu'il n'y a, bien évidemment, qu'un et un seul alignement satisfaisant les critères fixés.

Résultats Les résultats sont présentés dans la table 5.3.

Une séquence localisée sur un modèle de gène doit ensuite être signalée sur la page d'annotation de ce modèle. Cela permet de prévenir l'annotateur que ce modèle de gène correspond en fait à une séquence connue et publiée.

Tab. 5.3 – Intégration des connaissances biologiques dans l'annotation des génomes du projet Génolevures 3.

Les séquences publiées qui ne sont pas identifiées en tant que telles par l'annotation sont recherchées par alignement de type BLASTp dans l'ensemble des séquences protéiques prédites. Critères de validation de l'alignement : 90% de similarité sur la longueur de l'alignement, e-valeur maximale = 1.E-40, sauf pour KLLA dont les critères sont de 98% de similarité de séquence et une e-valeur de 0.01. (ZYRO : *Z. rouxii*, SAKL : *S. kluyveri*, KLTH : *K. thermotolerans*, KLLA : *K. lactis*). *K. lactis* faisant l'objet d'une réannotation, cette dernière doit tenir compte des annotations précédentes (5 402). La séquence étant mieux séquencée, 5 265 séquences publiées n'ont pas gardé le même nom lors de la réannotation et ont dû être recherchées par alignement : 5 204 séquences ont ainsi été localisées sur la nouvelle version de séquence ADN, 58 ont disparu de la nouvelle version, 3 n'ont pas été retenu car trop éloignées de la nouvelle séquence.

	ZYRO	SAKL	KLTH	KLLA
séquences publiées	73	116	5	5 402
séquences identifiées	29	56	0	137
séquences non identifiées	44	60	5	5 265
analyse BLASTp				
séquences localisées	44	60	5	5 204
séquences sans hit	0	0	0	58
séquences non retenues	0	0	0	3

5.3 Conclusion et perspectives

Les règles de cohérence appliquées jusqu'à présent sur les données du projet Génolevures 2 nous ont permis de détecter des erreurs dans les annotations réalisées. Ces erreurs concernent divers aspects d'une annotation : l'architecture des gènes codant les protéines, la syntaxe du commentaire, le défaut de prédiction de certains ORF. En détectant ces erreurs et en les corrigeant, nous améliorons ainsi la cohérence des annotations de ces génomes de levures. Notre méthode se révèle ainsi efficace pour l'amélioration de la qualité d'une annotation génomique. Intégrée au processus d'annotation, cette méthode permet de rectifier l'annotation au fur et à mesure de son avancée. Ainsi, les analyses basées sur l'annotation génomique sont plus justes. Ces règles sont à l'heure actuelle également appliquées sur les annotations en cours des levures étudiées par le projet Génolevures 3 (*Z. rouxii*, *S. kluyveri* et *K. thermotolerans*).

Perspectives

Pour autant, ce travail doit être poursuivi. En effet, nous voyons plusieurs perspectives à court et moyen termes.

À court terme, il faudrait automatiser l'application des règles liées à la structure et au non-chevauchement des gènes au système d'annotation semi-automatique MAGUS utilisé pour le projet Génolevures 3. Lorsque l'annotateur soumet une validation d'annotation pour un modèle de gène, ces règles devraient alors être appliquées afin de signaler une éventuelle erreur mais non empêcher la soumission. En effet, la recherche des exons et intron(s) d'un gène, par exemple, peut nécessiter des analyses complémentaires et qui demandent plus de temps. De plus, comme nous l'avons vu, certains cas, tels que le chevauchement de séquences codantes peuvent réellement exister.

Il faudrait également prendre en compte de façon plus appuyée les connaissances biologiques afin d'évoluer, dans ce cas, vers une annotation semi-automatique. De même, les molécules d'ARN devraient être recherchés et situés sur le génome à annoter avant de commencer l'annotation des modèles de gènes.

Ensuite, il faudrait implémenter les règles concernant la conservation des voies métaboliques et des interactions protéique-protéique. Par exemple, *C. glabrata* est connu pour ses difficultés à croître en présence de galactose au lieu de glucose [Nakase et al., 1998]. En effet, l'annotation génomique de cette levure a révélé que *C. glabrata* a subi une perte corrélée de 13 gènes sur les 30 impliqués dans la voie métabolique du galactose (d'après le KEGG) [Iragne et al., 2007].

À moyen terme, les données utilisées comme support pour les règles pourraient être enrichies, telles les clusters de gènes orthologues COG pour les règles basées sur l'inférence (*e.g.* conservation des voies métaboliques, conservation des fonctions essentielles), ou bien les complexes protéiques et nucléoprotéiques bien décrits dans la littérature.

Chapitre 6

Validation expérimentale de l'annotation

Sommaire

L'annotation génomique d'un génome, réalisée d'après des méthodes prédictives, est un premier pas vers la compréhension du fonctionnement d'un organisme. Elle permet de cibler les expériences *in vivo* à mener par la suite qui, en retour, confirmeront ou non

les prédictions de cette annotation. Nous avons choisi d'appliquer notre méthode de vérification à l'annotation de la levure *Yarrowia lipolytica* effectuée par le consortium Génolevures en étudiant son complexome. Cette étude procède selon deux axes. D'une part, l'identification expérimentale des protéines permet de confirmer les prédictions obtenues. D'autre part, l'identification expérimentale des interactions protéine-protéine permet de valider l'induction *in silico* de ces IPP à partir de celles de *S. cerevisiae*.

Nous avons choisi la méthode expérimentale de l'électrophorèse bi-dimensionnelle Blue Native-SDS (BN/SDS PAGE)[Schägger et al., 1996, Camacho-Carvajal et al., 2004] qui permet d'identifier les membres des complexes protéiques à l'échelle de la cellule en condition non dénaturante.

Nous présentons dans un premier temps les objectifs de cette expérimentation biologique, ainsi que les raisons des divers choix effectués (espèce étudiée, éléments étudiés, technique employée). Puis nous décrivons la méthode BN/SDS PAGE et les levures utilisées. Nous présentons ensuite les résultats expérimentaux obtenus concernant l'identification des protéines, complexées ou non, et les comparons à ceux obtenus par prédiction.

6.1 Objectifs et contexte

Les expériences ont été réalisées au Pôle Protéomique de la Plateforme Génomique Fonctionnelle de Bordeaux, à l'Université Victor Ségalen Bordeaux 2. Ce Pôle est dirigé par le Professeur Marc Bonneu, directeur de l'École Supérieure des Techniques des Biomolécules de Bordeaux.

6.1.1 Objectifs

Notre étude des complexes protéiques chez *Y. lipolytica* par une méthode expérimentale passe par la réalisation de quatre objectifs.

Le premier objectif est la validité de la technique expérimentale utilisée. Pour cela, nous avons choisi d'utiliser *S. cerevisiae* en tant que contrôle interne. Nous devons retrouver par notre technique certaines des interactions protéiques identifiées auparavant lors des nombreuses études expérimentales réalisées à grande échelle [Uetz et al., 2000, Ito et al., 2001, Gavin et al., 2002, Ho et al., 2002].

L'objectif suivant est de valider l'annotation génomique réalisée par le consortium Génolevures. Cette annotation a été faite manuellement par un ensemble d'annotateurs et vérifiée selon les règles de cohérence présentées dans le chapitre précédent. Cette annotation résulte d'une stratégie d'annotation dont certains paramètres ont été fixés par les membres du consortium. Ces paramètres sont, par exemple, la longueur minimale de détection d'un cadre de lecture ouvert (fixée à 80 nt pour la prédiction chez *Y. lipolytica*), les motifs d'épissage 5', 3' et point de branchement, les distances entre ces sites d'épissage (cf. § 5.1.2.2 chap5 p. 98). Pour *Y. lipolytica*, l'annotation dont nous disposions reposait à 96% sur des prédictions (seuls 234 gènes ou séquences étaient identifiés avant l'annotation du génome complet par le consortium Génolevures). La mise en évidence expérimentalement de protéines prédites permet de confirmer ainsi la validité de la stratégie d'annotation adoptée.

Le troisième objectif est de valider l'induction des règles de cohérence par la mise en évidence des complexes protéiques. En effet, comme nous l'avons vu § 4.3.2.2 p. 81, nous pouvons induire les interactions proteine-protéine chez les levures étudiées par le projet Génolevures, dont *Y. lipolytica*, à partir de celles identifiées expérimentalement chez *S. cerevisiae* [Uetz et al., 2000, Ito et al., 2001, Gavin et al., 2002, Ho et al., 2002]. Nous pouvons ainsi induire les complexes potentiels chez ces levures. Les différences observées entre les complexes prédits et ceux de *S. cerevisiae* révèlent seulement les complexes présents chez *S. cerevisiae* et absents des autres levures nouvellement séquencées. La mise en évidence expérimentale des complexes chez *Y. lipolytica* permet de confirmer d'une part, la prédiction des IPP pour *Y. lipolytica*, et d'autre part, d'identifier de nouvelles interactions. Le choix adéquat des espèces de levures utilisées pour cette expérimentation biologique est prépondérant à ce niveau-là (cf. § 6.1.2 ci-après).

Le dernier objectif est d'obtenir de nouvelles données expérimentales pour une espèce nouvellement séquencée. L'apport de ces données enrichit les données expérimentales existantes pour les levures étudiées au cours du projet Génolevures. Ainsi les comparaisons entre génomes s'appuient sur plus de données fiables. La technique expérimentale choisie doit alors permettre l'obtention de données protéomiques et complexomiques.

6.1.2 Choix des levures

Nous avons choisi d'étudier deux levures : *S. cerevisiae* et *Y. lipolytica*.

La validation de notre technique doit être réalisée avec un organisme dont le maximun d'interaction protéine-protéine est connu afin d'obtenir des résultats communs. Plusieurs études sur les IPP chez *S. cerevisiae* ont été réalisées à grande échelle par des techniques complémentaires [Uetz et al., 2000, Ito et al., 2001, Gavin et al., 2002, Ho et al., 2002]. Nous devons ainsi retrouver par notre technique au moins une partie des protéines complexées identifiées auparavant pour cette levure.

Puis nous avons choisi d'étudier expérimentalement *Y. lipolytica*, une des levures annotées au cours de la phase 2 du projet Génolevures, pour trois raisons. Tout d'abord, les connaissances expérimentales de *Y. lipolytica* concernaient principalement ses capacités d'expression et de sécrétion de protéines [Barth and Gaillardin, 1996, Beckerich et al., 1998, Nicaud et al., 2002]. En particulier, une étude à grande échelle faisait alors défaut. Ensuite, *Y. lipolytica* est la levure la plus distante de *S. cerevisiae* parmi les levures du projet Génolevures 2. Les informations expérimentales obtenues nous permettent ainsi d'encadrer la comparaison des levures hémiascomycètes. Enfin, *Y. lipolytica* est une levure dimorphique : elle adapte son métabolisme et sa physiologie selon son environnement de croissance. En présence d'un milieu de croissance enrichi en acides gras, son métabolisme de dégradation de ces acides est renforcé, en particulier au niveau des peroxysomes qui se développent. Cela se traduit par une modification de l'expression des protéines, par exemple par une sur-expression des gènes codant les enzymes, dont les lipases, intervenant dans les voies métaboliques de la β-oxydation des acides gras. En cultivant ainsi *Y. lipolytica* sur un milieu de culture enrichi ou non en acides gras, nous pouvons observer ces différences d'expression géniques en comparant la présence des protéines et complexes protéiques impliqués.

6.1.3 Intérêts des complexes protéiques

L'étude du complexome d'un organisme présente un intérêt double. D'une part, une cellule ou un organisme vivant doit être considéré comme une seule et même entité fonctionnelle. Ce sont tous ses composants qui, en interagissant ensemble, le font vivre. Ainsi identifier les interactions protéine-protéine existantes au sein d'une cellule permet d'approfondir les connaissances sur le fonctionnement de cette cellule ou organisme unicellulaire. De même, à un niveau supérieur, connaître les relations entre cellules d'un organisme multi-cellulaire permet d'approfondir les connaissances sur le fonctionnement de cet organisme. Il a été ainsi démontré que certaines protéines impliquées dans une même voie métabolique, se complexaient. Ainsi, des études ont permis de proposer une fonction à des protéines inconnues mais participant à un complexe de fonction connue.

D'autre part, la comparaison des complexes protéiques entre diverses espèces proches d'un point de vue phylogénique, permet d'obtenir quelques indices sur l'évolution de ces espèces. Comme nous l'avons vu dans le chapitre 2, § 2.4 p. 28, les complexes protéiques interviennent dans de nombreux processus biologiques (transport, voies métaboliques, voies de signalisation,...) dont des processus essentiels pour la cellule tels que la transcription ou la traduction. La pression de sélection exercée sur les complexes protéiques mis en jeu est par conséquent très forte. Les interactions doivent être conservées afin que la fonction soit préservée. De plus, la formation de complexes protéiques spécifiques chez une espèce révèle les particularités physiologiques de celle-ci. Ainsi l'adaptation des complexes reflète l'évolution des espèces et leur spéciation.

6.1.4 Choix de la méthode expérimentale

Nous avons vu au chapitre 2 § 2.4.4 p. 31, les diverses méthodes expérimentales actuelles couramment utilisées pour la mise en évidence des interactions protéine-protéine. Pour rappel, ces méthodes sont la technique du double-hybride, la purification en tandem (ou TAP-tag) et l'électrophorèse bi-dimensionnelle en gels bleu natif et SDS (BN/SDS PAGE). Nous avons également présenté les avantages et inconvénients de chacune de ces méthodes au chapitre 3 § 3.1.5.1 p.52.

Le contexte dans lequel s'est déroulée cette thèse nous a amené naturellement à choisir la méthode de l'électrophorèse 2D BN/SDS pour les raisons suivantes.

Tout d'abord, nous devions mettre en place une stratégie expérimentale qui permettrait d'obtenir des résultats peu de temps après l'obtention de l'annotation génomique de l'espèce. Des trois méthodes, la BN/SDS PAGE est la seule à satisfaire cette contrainte, car elle ne suppose pas la connaissance préalable des séquences géniques avant de commencer véritablement l'expérimentation. Ainsi, cette méthode peut être appliquée en parallèle d'un projet d'annotation génomique ; et ceci, pour n'importe quelle espèce.

Ensuite, en théorie, cette méthode permet l'observation peu intrusive de l'ensemble des complexes protéiques à l'échelle de la cellule entière, dans des conditions plus respectueuses des conditions normales de vie des cellules.

6.2 Matériels et méthodes

6.2.1 Cultures cellulaires

La souche *Y. lipolytica* utilisée est la souche E150 (ou CBLIB122), de phénotype Leu- His- Ura- Xpr- MatB, séquencée et annotée au cours des projets Génolevures 1 [Casaregola et al., 2000] et Génolevures 2 [Dujon et al., 2004].

La souche *S. cerevisiae* utilisée est la souche de laboratoire transformée BY 4742 Ura- His- Leu- + [GFP-énolase + Ura+] sur plasmide (gènes pour la synthèse des acides aminés uracile, histidine et leucine déficients, apport d'un gène ura+ par plasmide).

Composition des milieux de croissance utilisés :

- YPD liquide : 1% w/v peptone, 1% w/v extrait de levures, 2% w/v glucose,
- YDP solide : 1% w/v peptone, 1% w/v extrait de levures, 2% w/v glucose, 2% agar,
- YP_2DH_5 liquide : 1% w/v peptone, 1% w/v extrait de levures, 2% w/v glucose, 5% w/v huile d'olive, tamponné avec du NaH_2PO_4.

Les deux souches sont conservées sur milieu solide YPD sur boîte de Pétri au réfrigérateur à 7°C et sont repiquées tous les 15 jours.

6.2.1.1 Préculture

Les souches sont mises en préculture dans 80 mL d'YPD liquide sous agitation (120 rpm) en condition aérobie avec les particularités suivantes :

- pour *S. cerevisiae* : à 30°C pendant 16 h,
- pour *Y. lipolytica* : à 28°C pendant 12 h.

6.2.1.2 Culture

Les cultures cellulaires se font dans des flacons de 2 L contenant 500 mL de milieu de croissance liquide pour 16 mL de préculture, sous agitation (120 rpm), en condition aérobie, avec les particularités suivantes :

- pour *S. cerevisiae* : milieu YPD, à 30°C, pendant 16 h (fin de la phase exponentielle de croissance cellulaire),
- pour *Y. lipolytica* : à 28°C,
  - métabolisme normal : milieu YPD, pendant 12 h (fin de la phase exponentielle de croissance cellulaire),
  - métabolisme lipidique activé : milieu YP_2DH_5, pendant 76h (niveau maximal de synthèse des lipases [Nicaud et al., 2002]).

6.2.2 Préparation des échantillons

Deux types de fractions sont obtenus : la fraction contenant les protéines et complexes protéiques cytoplasmiques, et celle enrichie en protéines et complexes membranaires. Le mode de récolte des cellules est identique quel que soit le milieu de culture et la souche. Les étapes supplémentaires signalées comme telles concernent uniquement *Y. lipolytica* cultivée en milieu YP_2DH_5.

Pour *Y. lipolytica* cultivée en présence de YP_2DH_5, avant récolte, le maximum d'huile surnageante est enlevé à la pipette à température ambiante.

Les cellules sont récoltées par centrifugation pendant 7 min à 4℃ à 4 640 × g. À partir de cet instant, la procédure d'extraction des protéines se déroule à 4℃ (appareils réfrigérés ou bac de glace) sauf précision contraire afin de préserver au maximum les complexes protéiques. Le surnageant est jeté. Les culots sont rincés dans de l'eau pure. Un second rinçage est effectué pour *Y. lipolytica* cultivée sur huile.

Les cellules sont récoltées par centrifugation pendant 7 min à 4 640 × g. Les cellules sont resuspendues dans le tampon d'extraction contenant 750 mM d'acide 6-amino-n-caproïque (Avocado) et 50 mM de Tris (Amersham [1]). Elles sont récoltées par centrifugation pendant 7 min à 4 640 × g. Une seconde centrifugation est effectuée pour *Y. lipolytica* cultivée sur huile.

Les cellules sont resuspendues avec du tampon d'extraction (1 w/v au maximum), 1 mM PMSF (inhibiteur de protéases, Avocado), et un cocktail d'inhibiteurs de protéases spécifiques des levures et champignons (à 1 mL/20 g levures entières, Sigma).

Les cellules sont cassées par deux passages au travers de la cellule de presse French à 2 kbars. Les cellules non cassées et les morceaux de parois cellulaires sont enlevés par centrifugation pendant 20 min à 12 100 rpm.

La DNase (DNase bovine, Sigma) à 2 mg/mL est ajoutée pour atteindre 1/10 du volume final. Le traitement à la DNase est réalisé à 25℃ pendant 1 h 20. En effet, représentant une molécule de taille importante, l'ADN gène ensuite la purification et la migration des protéines, complexées ou non. Du fait de la robustesse des nucléotides, l'ADN doit être clivé par une enzyme spécifique, la DNase. L'échantillon est centrifugé pendant 45 min à 100 000 × g à 4℃. Le surnageant et le culot correspondent respectivement aux protéines cytoplasmiques et aux protéines membranaires.

Les protéines cytoplasmiques sont filtrées sur membrane Miracloth (Calbiochem) pour enlever les corps gras. Cette fraction est utilisée directement ou stockée à -20℃.

Trois cents μg de fraction cytoplasmique sont déposés sur une colonne HiTrap de dessalage (Amersham) équilibrée au tampon d'extraction. Les fractions contenant les protéines sont utilisées directement ou stockées à -20℃.

La fraction cytoplasmique (300 μg) peut également être concentrée par ultrafiltration à centrifugation Vivaspin 500, à 100 kDa (Vivascience). Trois cents μL sont déposés dans l'unité d'ultrafiltration et soumis à une centrifugation de 3 000 × g de façon à obtenir 50 μL de retentât. Les protéines sont concentrées 6 fois par rapport au dépôt initial. Cette étape de concentration peut être utilisée comme dessalage.

Le culot membranaire est resuspendu dans du tampon d'extraction avec ajout de 1 mM PMSF, toujours à 4℃. L'échantillon est recassé en un seul passage à la cellule de presse French à 2 kbars. Lors de la première lyse cellulaire, le cytoplasme des cellules et le contenu des organelles et vésicules sont libérés et récupérés dans la fraction cytoplasmique. Mais les membranes se reforment rapidement, de par leur structure hydrophobe, enfermant ainsi une

[1] Amersham fait désormais partie de GE Healthcare.

partie des protéines et complexes membranaires à l'intérieur, hors d'atteinte des réactifs. En refractionnant les membranes dans un volume plus grand, nous supposons que les morceaux de membrane se répartissent uniformément dans ce volume et se retrouvent plus distants les uns des autres. Les complexes protéiques membranaires représentant une structure volumineuse, ces morceaux ne pourraient plus se refermer (du moins, avec plus de difficulté) et emprisonner ainsi les complexes. Les complexes seraient alors plus accessibles aux produits chimiques. Le lysat membranaire est centrifugé pendant 45 min à 100 000 × g. Le surnageant est jeté. Le culot est repris dans du tampon d'extraction additionné de dodecyl-β-D-maltoside (Sigma) à une concentration finale de 2% w/v. Ce réactif solubilise les protéines de la membrane interne. L'échantillon est conservé 15 min à 4°C. Puis il est centrifugé pendant 45 min à 100 000 × g. Le surnageant contient les protéines membranaires. Il peut être utilisé directement ou stocké à -20°C.

Les échantillons protéiques cytosoliques et membranaires sont prêts pour être chargés sur le gel de première dimension.

6.2.3 Électrophorèse BN/SDS

L'électrophorèse BN/SDS reprend les principes élaborés par Schägger dès 1987 [Schägger and von Jagow, 1987] qui n'a cessé de la perfectionner [Schägger and von Jagow, 1991, Schägger, 1995, Schägger et al., 1996, Schägger, 2006, Wittig et al., 2006]. Elle se déroule en deux étapes principales :

- l'électrophorèse en première dimension en bleu natif (électrophorèse 1D BN) : les protéines non complexées et les complexes protéiques migrent en condition non dénaturante selon leur poids moléculaire et leur forme,
- l'électrophorèse en seconde dimension en SDS (électrophorèse 2D SDS) : les complexes sont dissociés, les protéines migrent selon leur poids moléculaire.

6.2.3.1 Électrophorèse 1D BN

Les gels de séparation pour la première dimension ont été utilisés selon divers gradients linéaires de polyacrylamide : 4-18%, 3-10%, 9-18%, 8-14%. Leur réalisation suit la méthode de Schägger. La composition des gels de polyacrylamide 1D BN se trouve dans la table 6.1.

Les plaques utilisées pour le gel 1D ont une taille de 22 cm × 16.5 cm × 0.1 cm. Le gel de séparation est coulé verticalement entre les plaques à l'aide du mélangeur des deux solutions (par exemple les solutions à 4 et 18% pour un gel de gradient 4-18%), selon le gradient de polyacrylamide voulu, le bas du gel ayant le plus fort pourcentage de polyacrylamide. Un filet de butanol est rajouté sur le haut du gel afin qu'il polymérise sans ménisque ni aspérité et sans se rétracter. Après polymérisation, le butanol est enlevé avec rinçage à l'eau.

Le gel de concentration à 3% de polyacrylamide est ensuite coulé au-dessus du gel de séparation, en disposant le peigne (le peigne utilisé pour les plaques dispose de 12 puits). Lors de la migration, ce gel permet aux protéines de pénétrer en même temps dans le gel. Après polymérisation, le gel est placé au réfrigérateur à 7°C, dans la cuve à électrophorèse verticale, en conservant le peigne en place (toujours pour éviter que le gel ne se déshydrate) au contact

TAB. 6.1 – Composition des gels de polyacrylamide 1D BN.

[3%] indique la composition pour le gel de concentration à 3%. Le tampon 3× contient 1.5 M acide 6-amino-caproïque et 150.10^{-3} M Tris. Le TEMED et l'APS doivent être rajoutés juste avant de couler le gel car ils déclenchent la réaction de polymérisation du gel.

% du gel	3%	4%	8%	9%	10%	14%	18%	[3%]
acrylamide 40% (μL)	1800	1500	3000	3375	3750	5250	6750	900
bis-acrylamide 2% (μL)	1120	937.5	1875	2110	2354	3280	4222	560
Tp 3× (μL)	8000	5000	5000	5000	5000	5000	5000	4000
glycérol 87% (g)	\	\	3	3	3	3	3	\
eau (μL)	1340	7560	2125	1515	906	\	\	6520
APS 10% w/v (μL)	120 / 72	120	60	60	60	60	60	32
TEMED (μL)	12 / 7.2	12	6	6	6	6	6	3.2

de l'air, jusqu'à utilisation. le gel est placé avec les tampons anode et cathode dans leur bac respectif. Les tampons anode et cathode contiennent 50 mM Tris, 75 mM glycine (Amersham PlusOne). Le tampon cathode contient en plus du Bleu Serva G (Serva, Heidelberg, Germany) à 0,002% w/v. Avant d'être chargé dans le puits, 1 μL de tampon de charge (500 mM acide 6-amino-n-caproïque, 5% w/v Bleu Serva G) est ajouté à l'échantillon (d'un volume de 12 à 25 μL). Le Bleu Serva G amène des charges négatives sur les protéines de façon uniforme sur leur séquence (au moins pour la partie accessible).

Chaque échantillon est déposé en double : un exemplaire est utilisé pour la seconde électrophorèse en SDS, l'autre sert de contrôle. Le gel est mis sous tension à 1 W (100 V, 30 mA) pendant 36 h à 7℃. Sous tension, la température du gel s'élève : afin de ne pas abîmer les protéines, le gel est placé dans une armoire réfrigérée. Les protéines thyroglobuline (669 kDa) et serum-albumine bovine (BSA, 66 kDa) sont utilisées comme marqueurs de poids moléculaires standards (Sigma-Aldrich) pour l'analyse BN PAGE. Les protéines et complexes protéiques, chargés négativement, migrent ainsi selon leur poids moléculaire (existence d'une corrélation positive entre la taille et la charge négative de la protéine ou du complexe) et leur forme (cf. fig. 2.19 étape A, p. 35).

La suite des expériences est réalisée à température ambiante. Après la migration, le gel est démoulé. L'ensemble des pistes contrôles est coloré au Bleu de Coomassie G250 0.125% w/v (Bleu/éthanol/eau/acide acétique 0.125 : 50 : 40 :10) pendant 20 min sous agitation légère. Puis le gel est décoloré à l'acide acétique (éthanol/eau/ac. acétique 25 : 67 : 8) afin de révéler seulement les bandes contenant les protéines.

Cette coloration permet de vérifier d'une part la bonne migration des complexes et protéines (visualisation de bandes fines et droites), et d'autre part, la présence de matériel protéique en quantité suffisante pour faire la seconde électrophorèse (appréciation à l'œil nu). Les bandes présentant suffisamment de matériel pour la seconde électrophorèse sont découpées individuellement.

6.2.3.2 Électrophorèse 2D SDS

Une étape d'équilibration des bandes issues du gel 1D est nécessaire avant l'électrophorèse 2D SDS proprement dite afin de rompre les liaisons disulfures entre les acides aminés cystéines et dissocier ainsi les complexes. Les protéines seront ainsi réduites. Le SDS, chargé négativement, va ainsi se répartir le long de la protéine, masquant la charge naturelle de la protéine. Les protéines, chargées alors négativement proportionnellement à leur taille, migreront en fonction de leur charge, *i.e.* leur taille.

Chaque piste du gel 1D est découpée, placée dans un tube et équilibrée 5 min avec le tampon d'équilibration 0.125 mM Tris/HCl pH 6.8 et 1% w/v SDS (Amersham) sous agitation légère. Puis la bande est trempée de nouveau pendant 13 min dans le tampon d'équilibration en présence de 100 mM de DTT (Proméga) sous agitation légère. Le DTT rompt par réduction les ponts disulfures, liaisons de forte intensité formées entre les groupements thiols de deux cystéines, et intervenant dans la structure 2D, 3D et 4D de la protéine (cf.2§ 2.2.3 p. 19). La bande est ensuite baignée pendant 13 min dans du tampon d'équilibration en présence de 55 mM de iodoacétamide (Sigma-Aldrich) sous légère agitation. Le iodoacétamide empêche les ponts disulfures de se reformer par alkylation (ajout d'un groupement alkyl sur le groupement thiol de la cystéine). L'étape d'équilibration de la bande se termine par un rinçage pendant 5 min avec du tampon d'équilibration sous agitation légère.

La bande issue du gel de première dimension est disposée en haut d'une des plaques pour le gel 2D, horizontalement, en lieu et place du gel de concentration, proche de l'un des bords. Après disposition des espaceurs, la seconde plaque vient se superposer à la première. Les plaques sont maintenues en place par des pinces. Il s'agit alors de vérifier l'absence de bulles d'air au niveau de la bande et la disposition de la bande. En effet, la compression de la bande entre les plaques déforme celle-ci, or la bande doit rester droite afin que les protéines migrent au maximum en restant dans leur "couloir" de migration. De même la présence de bulles d'air entre la bande et les plaques peut gêner la migration car modifie à cet endroit le passage du courant électrique. Les plaques sont mises en position verticale afin de couler le gel 2D SDS. La composition du gel se trouve table 6.2. Le gel de séparation à 10% de polyacrylamide est alors coulé. Pour des raisons identiques à celles évoquées avec le gel 1D BN, un filet de butanol est disposé au-dessus du gel de séparation. Après polymérisation, le gel de concentration à 4% de polyacrylamide est coulé à son tour, recouvrant la bande issue du gel 1D.

Le gel est ensuite placé dans la cuve d'électrophorèse (Amersham). La cuve utilisée peut accueillir jusqu'à 6 gels 2D. La cuve est remplie de tampon de migration 1× (le tampon 10× contient 430 mM glycine, 200 mM Tris et 26 mM SDS). Le gel est mis sous tension pour la nuit. Pour 6 gels, le réglage est de 500 V, 120 mA (il faut entre 12 et 20 mA par gel, sachant que le circuit électrique est en parallèle), 4 W (au maximum 17 W par gel). Les protéines migrent alors en fonction de leur poids moléculaire.

À l'issue de la migration, le gel est démoulé. Les protéines sont colorées en utilisant le kit de coloration Argent PROTSIL1 (Sigma-Aldrich) selon les recommandations du fabriquant.

TAB. 6.2 – Composition du gel de polyacrylamide 2D SDS.

Le tampon de séparation contient 1.5 M Tris/HCl pH=8.8, le tampon de concentration contient 0.5 M Tris/HCl pH=6.8. Le TEMED et l'APS doivent être rajoutés juste avant de couler le gel car ils déclenchent la réaction de polymérisation du gel.

	gel de séparation 10%	gel de concentration 4%
acrylamide 40% (μL)	8 000	800
bis-acrylamide 2% (μL)	4 260	426
Tp séparation / Tp concentration (μL)	8 160	2 040
eau (μL)	11 400	4 634
SDS 20% w/v (μL)	160	40
APS (μL)	160	60
TEMED (μL)	16	6

6.2.4 Identification des protéines par LC-MS/MS

L'identification des protéines se fait par LC-MS/MS. La chromatographie en phase liquide (LC) permet de séparer les différents peptides de l'échantillon. Ceux-ci sont ensuite analysés par spectrométrie de masse en tandem sur une trappe ionique. Les peptides sont d'abord ionisés puis analysés en fonction de leur rapport m/z. Cette analyse donne lieu à un spectre dit "MS". Les ions majoritaires peuvent être sélectionnés, isolés et fragmentés par collision avec de l'hélium à l'aide d'une tension radiofréquence correspondant à leur fréquence de résonance. Les ions fragments produits sont piégés. Ils sont à leur tour analysés en fonction de leur m/z pour donner lieu à un spectre dit "MS/MS". Ces spectres MS/MS expérimentaux sont comparés à un ensemble de spectres théoriques obtenus par digestion et fragmentation *in silico* de protéines de séquences connues. Plusieurs critères de validation sont alors appliqués afin de déterminer le ou les meilleurs candidats parmi les protéines de la banque de données. Les spectres sont suffisamment caractéristiques des fragments pour identifier une protéine à coup sûr à partir de deux peptides.

6.2.4.1 Préparation des échantillons

Les protéines colorées au nitrate d'argent sont excisées (excision au scalpel pour les spots de *Y. lipolytica*, à l'emporte-pièce pour ceux de *S. cerevisiae*) et placées dans des plaques de 96 puits Les spots de gel sont décolorés en utilisant le kit de coloration Argent PROTSIL2 (Sigma-Aldrich) selon les recommandations du fabriquant. Les spots sont rincés deux fois à l'eau ultra pure puis réduits avec l'ACN (Sigma R Chromasolv) pendant 10 min. Une fois l'ACN enlevé, les spots sont séchés à température ambiante, puis baignés dans une solution de trypsine (Sigma) à 10 ng/μL dans 50 mM de NH_4HCO_3 (Sigma). Les spots sont réhydratés à 4°C pendant 10 min et finalement incubés pendant la nuit à 37°C. La trypsine est une endoprotéase qui hydrolyse la liaison peptidique située en aval des lysines et arginines. Les

protéines sont ainsi hydrolysées en divers peptides.

Les échantillons sont ensuite incubés 15 min dans 50 mM NH_4HCO_3 à température ambiante. Le surnageant est collecté. La solution d'extraction H_20/ACN/HCOOH (47.5 : 47.5 : 5) (HCOOH : Prolabo Normapur 20)) est ajouté sur les spots pendant 15 min. Cette étape d'extraction des peptides du gel est répétée deux fois. Les surnageants sont rassemblés et concentrés par centrifugation sous vide jusqu'à un volume final de 25 μL. Les peptides digérés sont ensuite acidifiés avec 1.5 μL d'acide acétique 5% v/v (Prolabo Normapur 20) et stockés à -20℃. Les échantillons sont prêts à être analysés par LC-MS/MS.

6.2.4.2 Analyse LC-MS/MS

Chaque échantillon peptidique est analysé par nanoHPLC (LC Packings, Amsterdam, Pays-Bas) couplée à un spectromètre de masse nanospray LCQIT (ThermoFinnigan, San José, Canada).

Les échantillons de *Y. lipolytica* sont analysés comme suit. Dix μL de digestats peptidiques sont chargés sur une précolonne C18 PepMap™ de 300 μm di × 5 mm (LC Packings) (di : diamètre interne) avec un débit de 30 μL/min. Les peptides sont élués à partir de la précolonne sur une colonne analytique C18 PepMap™ 75 μm di × 15 cm (LC Packings) avec un gradient linéaire de 5-50% de solvant B en 30 min (solvant A = H_20/ACN/HCOOH 94.9 : 5 : 0.1, solvant B = H_20/ACN/HCOOH 19.9 : 80 : 0.1). Le débit de la séparation est de 200 nL/min. Le spectromètre de masse opère en mode positif pour étudier les ions positifs. La tension appliquée à l'aiguille de spray est de 2 kV et celle appliquée au capillaire de transfert de 46 V.

L'acquisition des données est réalisée par la méthode "dépendante des données", alternant un scan MS (m/z 300-2000) et trois scans MS/MS effectués sur les 3 ions les plus intenses du spectre MS qui précède. L'option "exclusion dynamique", qui permet d'exclure les m/z déjà fragmentés, est activée. Les spectres MS/MS sont acquis en utilisant une fenêtre d'isolation de 2 m/z, une énergie de collision relative de 35%, et une durée d'exclusion dynamique de 0.5 min.

Les échantillons de *S. cerevisiae* sont analysés de la même façon avec les modifications suivantes. Le gradient linéaire est de 5-40% en 35 min. Le voltage de l'aiguille est de 1.8 kV et celui du capillaire de 3 V. Le mode d'acquisition alterne cette fois-ci un scan MS, un "Zoomscan" permettant d'accéder à l'état de charge des ions et un spectre MS/MS.

6.2.4.3 Analyse des spectres

Les données sont analysées par l'algorithme SEQUEST à travers l'interface Bioworks 3.1 (ThermoFinnigan) avec la banque de données de l'espèce étudiée, *S. cerevisiae* (fichier FASTA provenant de SGD, 6 719 entrées) ou *Y. lipolytica* (banque indexée de 6 436 entrées provenant de l'annotation génomique réalisée par Génolevures et des séquences disponibles sur Swiss-Prot), plus un ensemble de séquences spécifiques d'éléments contaminants potentiels (dont les kératines humaines et animales).

Les fichiers de données au format DTA (contenant la liste des intensités et des ions de chaque spectre MS/MS) sont générés pour les spectres MS/MS qui respectent les contraintes

d'intensité minimale (fixée à 5×10^4) et de nombre suffisant d'ions (fixé à 15). La génération du fichier DTA permet également d'établir la moyenne de plusieurs spectres MS/MS (à partir de 20 scans) correspondant à un même ion précurseur avec une tolérance de 1.5 Da. Les spectres provenant d'ions précurseurs supérieurs à 3500 Da ou inférieurs à 600 Da sont rejetés. Les paramètres de recherche des algorithmes d'identification sont les suivants :

- tolérance de masse sur l'ion précurseur : 1.5 Da,
- tolérance de masse sur les ions fragmentés : 0.5 Da,
- calcul de la masse prenant en compte seulement les ions b et y,
- modifications post-traductionnelles considérées : oxydation des méthionines (+16 Da) et carbamidométhylation des cystéines (+57 Da),
- tolérance de 2 clivages trypsiques manqués,
- validation des peptides ayant un Xcorr[2] supérieur à 1.5 (si monochargé), 2 (si doublement chargé) et 2.5 (si triplement chargé),
- $-\Delta C_n \geqslant -0.1$.

Les identifications des protéines se basent toutes sur un minimum de deux peptides différents assignés à une protéine, sauf indication contraire.

Les peptides de *S. cerevisiae* sont analysés de la même façon avec les modifications suivantes. La moyenne des spectres MS/MS est établie à partir de 10 scans. Les seuils de Xcorr sont fixés à 1.9 (si monochargé), 2.2 (si doublement chargé), 3 (si triplement chargé).

6.2.5 Remarques

La méthode expérimentale a pour objectif d'identifier les complexes protéiques et les protéines impliquées dans ces complexes, dans des conditions de vie naturelles pour les cellules. Des précautions doivent être prises lors des manipulations du matériel biologique afin de minimiser la séparation des partenaires des complexes protéiques et la dégradation des protéines.

Ainsi, il faut éviter les chocs thermiques. La cellule de levure pousse en condition normale à 28°C. Faire passer instantanément une cellule de 28°C à 4°C permet d' 'endormir' la cellule et ralentir ainsi les mécanismes biologiques. De plus, lors de la lyse cellulaire, les enzymes de dégradation de protéines, appelées protéases, alors confinées dans des compartiments cellulaires particuliers tels que la vacuole ou les peroxysomes (cf. § 2.1.3.2 et 2.1.3.2 p. 16), se trouvent en contact avec les protéines et complexes protéiques. Elles vont alors dégrader les protéines. Là aussi, le froid diminue l'activité de ces enzymes.

De même une variation de pH trop importante de la solution dans laquelle se trouve l'extrait cellulaire, va dénaturer les structures 3D et 4D des protéines : les complexes protéiques sont alors dissociés.

Par ailleurs, il est conseillé d'enchaîner certaines étapes. Lors de la récolte puis de la lyse des cellules, il est néanmoins possible de congeler un culot de cellules entières avant la lyse. Dès que les cellules sont cassées à la presse French, il est recommandé d'ajouter rapidement le PMSF afin de limiter l'activité protéolytique des enzymes. À l'issue de l'ultracentrifugation, les fractions cytoplasmiques et membranaires peuvent être congelés à -20°C.

[2]Xcorr : score de corrélation *masse/charge* ou m/z.

De plus, lors de la réalisation des gels 2D, il est conseillé de minimiser le temps d'exposition à l'air libre de la bande du gel 1D même lorsqu'elle se trouve entre les deux plaques.

Des précautions doivent également être prises dans la réalisation de certaines solutions dont la composition et, par conséquent, l'action se dégradent au cours du temps, à la lumière, à la chaleur, ou au froid. Le tampon d'équilibration, par exemple, doit être réalisé lors de l'étape d'équilibration des bandes. Les expériences ont été réalisées avec les produits des marques citées. Nous ne garantissons pas le même résultat avec des produits équivalents chez d'autres marques.

6.3 Résultats et discussion

Les complexes protéiques cytoplasmiques de *S. cerevisiae* et *Y. lipolytica* (cultivée en condition normale et en condition enrichie en lipide) sont séparés par électrophorèse BN/SDS.

Les tables 6.3, 6.4 et 6.5 présentent respectivement les protéines en complexes multimériques, en complexes homodimériques et seules (ou non attribuées de façon certaine comme faisant partie d'un complexe) pour *S. cerevisiae*. L'ensemble des protéines identifiées est localisé sur les figures 6.1 et 6.2.

Les tables 6.6, 6.7, 6.8 et 6.9 présentent respectivement les protéines en complexes multimériques, en complexes homodimériques et seules (ou non attribuées de façon certaine comme faisant partie d'un complexe) pour *Y. lipolytica*. L'ensemble des protéines identifiées est localisée sur les figures 6.3, 6.4 et 6.5. Pour des raisons de commodité, j'ai laissé les fonctions en anglais.

L'application de l'électrophorèse BN/SDS sur *S. cerevisiae* m'a permis de mettre en évidence 50 protéines distinctes de façon certaine (identification basée sur au moins deux peptides différents) et une protéine de façon incertaine (un seul peptide caractéristique) dont 26 sont complexées dans 9 complexes hétéromultimériques et 3 complexes homodimériques. Huit interactions identifiées entre partenaires ont déjà été caractérisées expérimentalement par d'autres techniques (double hybride, purification en tandem). Les banques de données, par exemple IntAct et BioGrid[3], centralisent les interactions protéine-protéine et les sources de leurs publications.

Chez *Y. lipolytica*, j'ai identifié 114 protéines prédites *in silico* (avec au moins deux peptides différents) dont 4 annotées comme ne présentant aucune similarité avec une séquence connue à ce jour. Douze protéines sont identifiées sur la base d'un seul peptide. Par ailleurs, j'ai mis en évidence 9 complexes hétéromultimériques et 7 complexes homomultimériques chez cette levure. L'étude de *Y. lipolytica* dans deux conditions de croissance m'a permis d'observer des profils d'expression protéiques différents.

6.3.1 La technique de l'électrophorèse BN/SDS

Lors de la première dimension, chacun des gradients de polyacrylamide permet d'optimiser la séparation et la focalisation des complexes (et des protéines non complexées) pour une

[3]bioGrid : http ://www.thebiogrid.org/index.php.

FIG. 6.1 – Gel 2D de *S. cerevisiae* (extrait cytoplasmique).

Gel 2D de polyacrylamide 10% d'un échantillon cytoplasmique (12.5 μL) de *S. cerevisiae* après migration sur un gel 1D de polyacrylamide 4-18%. Coloration Argent.

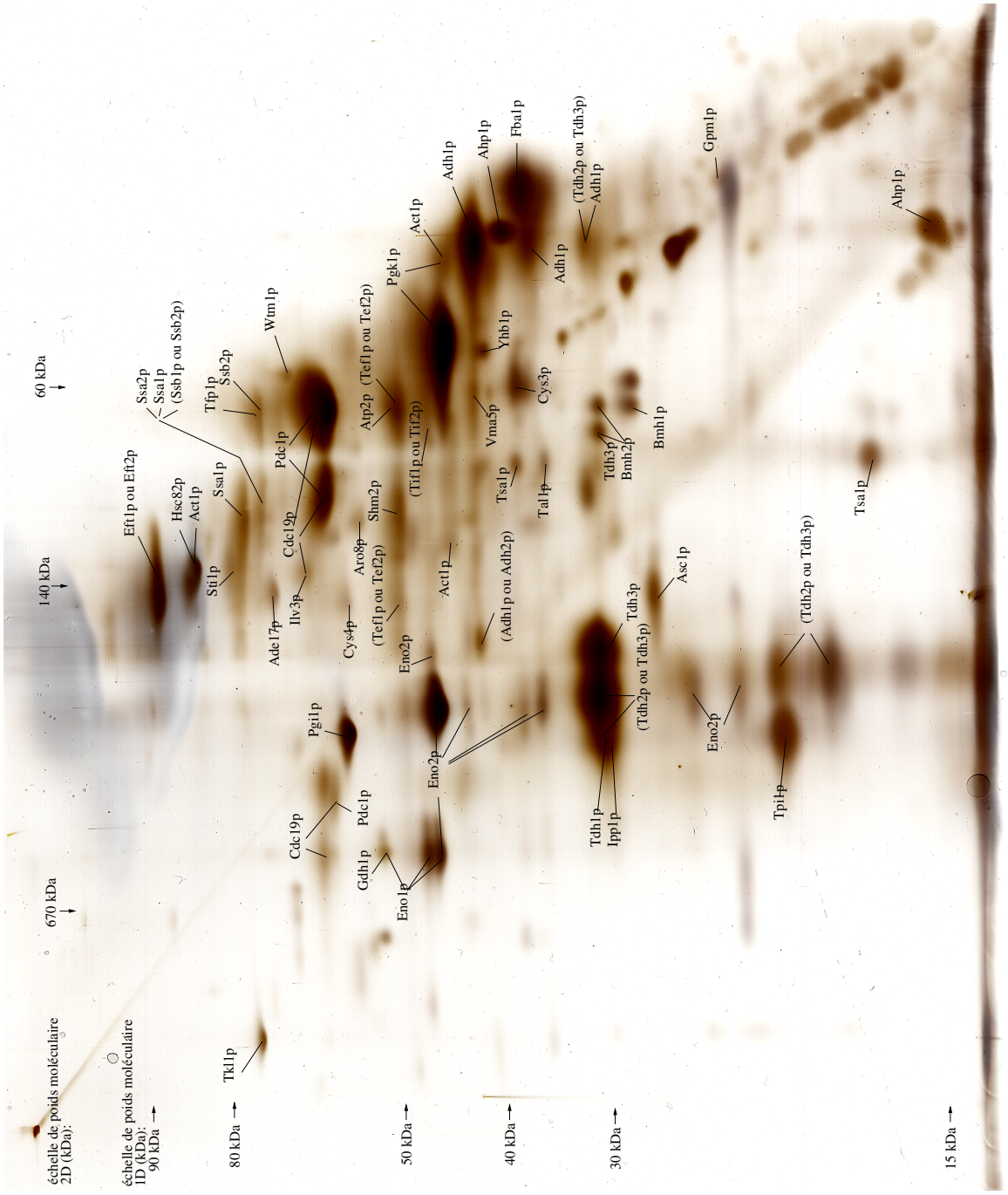


FIG. 6.2 – Gel 2D de *S. cerevisiae* (extrait cytoplasmique concentré).

Gel 2D de polyacrylamide 10% d'un extrait cytoplasmique concentré par Vivaspin (13 μL) de *S. cerevisiae* après migration sur un gel 1D de polyacrylamide 4-18%. Coloration Argent.

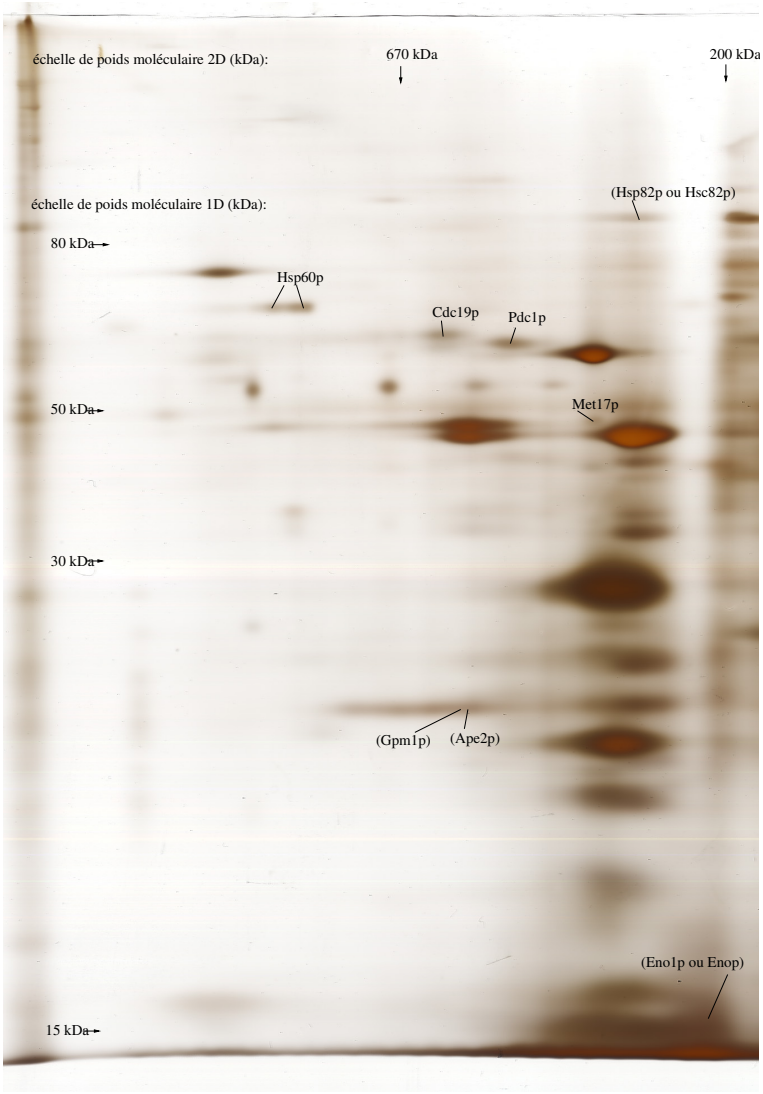


FIG. 6.3 – Gel 2D de *Y. lipolytica* (croissance sur huile, extrait cytoplasmique concentré)

Gel 2D de polyacrylamide 10% d'un échantillon cytoplasmique (13 μL) de *Y. lipolytica* (croissance sur milieu enrichi en huile) après migration sur un gel 1D de polyacrylamide 4-18%. Coloration Argent.

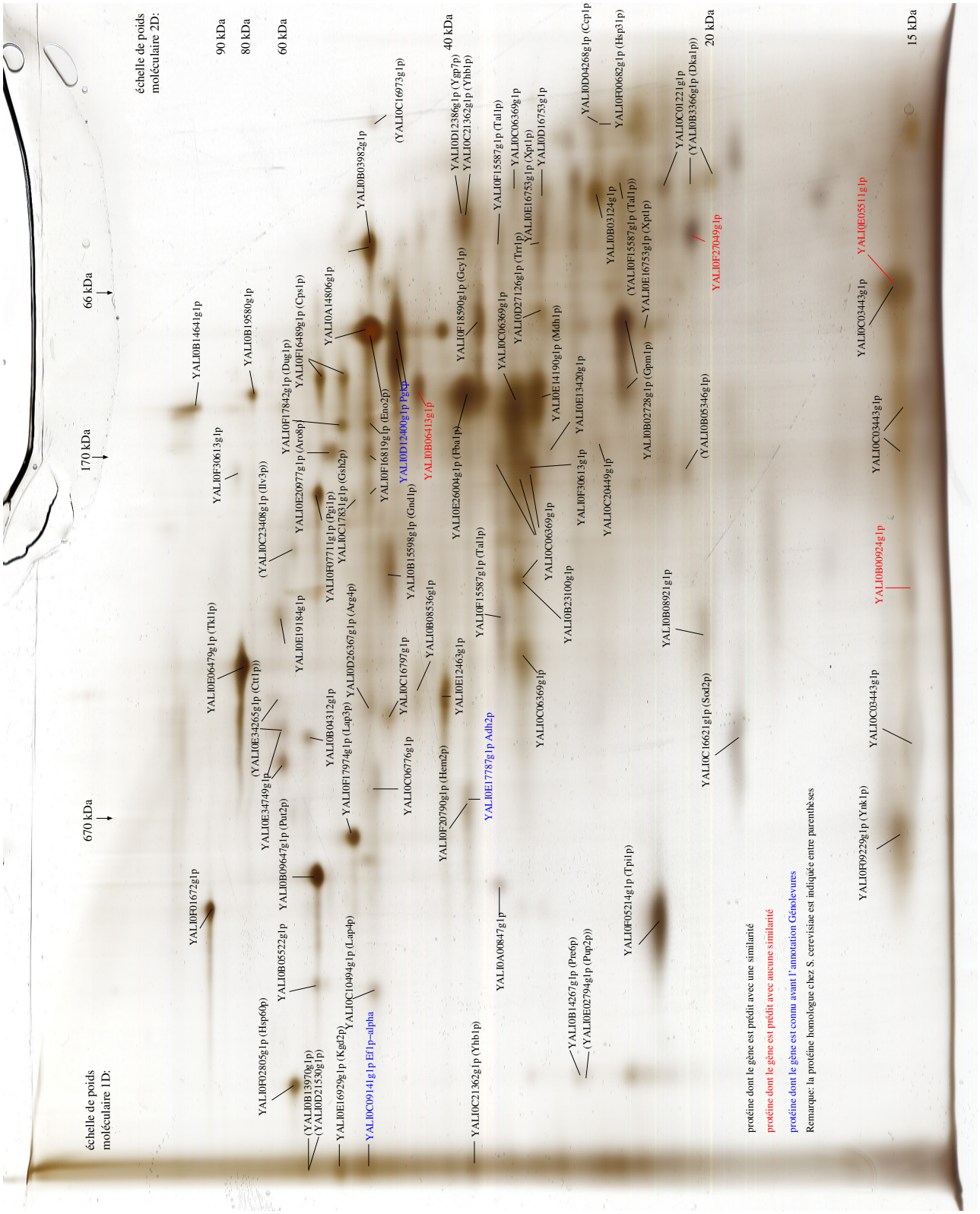


FIG. 6.4 – Gel 2D de *Y. lipolytica* (extrait cytoplasmique concentré)

Gel 2D de polyacrylamide 10% d'un échantillon cytoplasmique (13 μL) de *Y. lipolytica* (croissance sur milieu normal) après migration sur un gel 1D de polyacrylamide 4-18%. Coloration Argent.

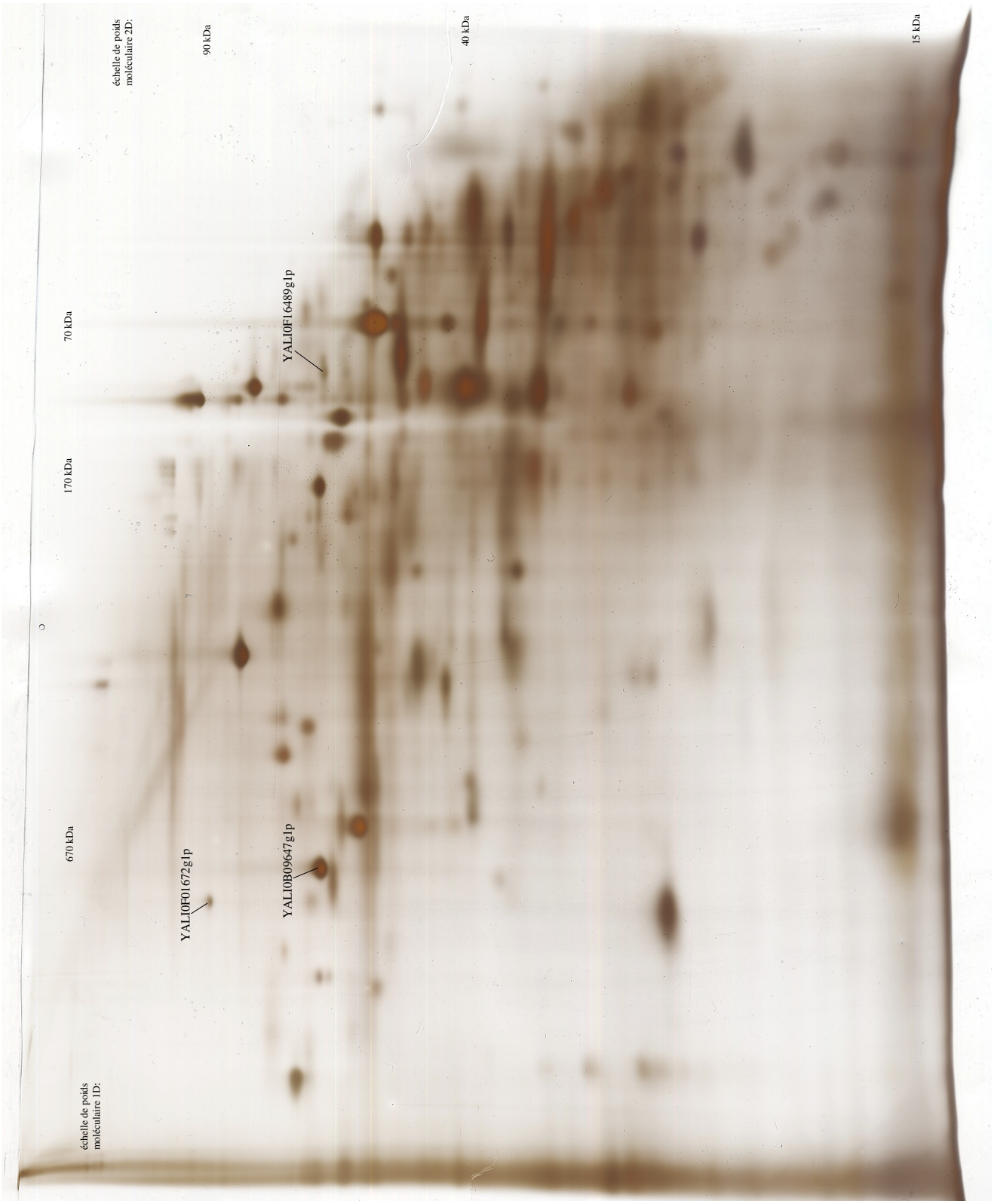
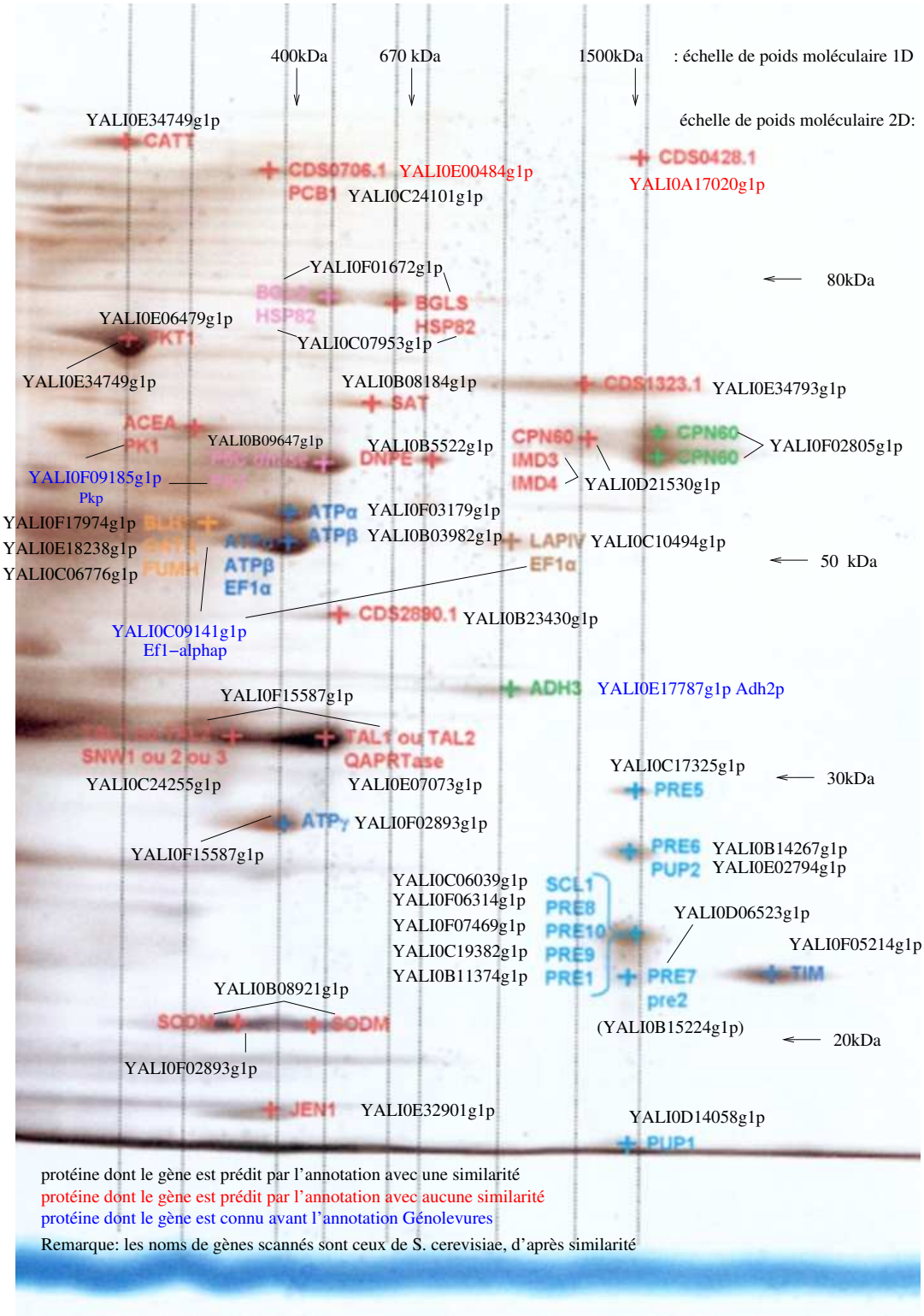


FIG. 6.5 – Gel 2D de *Y. lipolytica* (extrait cytoplasmique).
Gel 2D de polyacrylamide 10% d'un échantillon cytoplasmique (15 μL) de *Y. lipolytica* (croissance sur milieu normal) après migration sur un gel 1D de polyacrylamide 4-18%. Coloration Argent.



taille donnée. Lors d'une électrophorèse, les protéines migrent selon leur taille. Les complexes sont d'autant plus ralentis dans leur migration que la densité de maillage du gel est élevée. L'utilisation d'un gel avec un gradient de 8-12% w/v de polyacrylamide permet de séparer les complexes correspondant à la gamme de poids moléculaire donnée, sur une plus grande distance que lors de leur migration sur un gel de 4-18% w/v.

Les protéines appartenant à un même complexe co-migrent lors de la première dimension. Lors de la seconde dimension, les protéines étant dénaturées, elles se séparent et migrent chacune selon son poids moléculaire, de façon alignée et avec la même forme.

Par exemple, chez *S. cerevisiae*, Bmh1p (Yer177wp) et Bmh2p (Ydr099wp), qui forment un hétérodimère impliqué dans la voie de signalisation cellulaire [Chaudhri et al., 2003], migrent de façon alignée et l'une juste au-dessous de l'autre (cf. fig. 6.1).

6.3.2 Validité et limites de la méthode

Le laboratoire où j'ai réalisé ces expériences, appliquait cette technique en parallèle chez le colibacille *E. coli*. Les résultats obtenus plus rapidement chez *E. coli* [Lasserre et al., 2006] nous ont fortement encouragés à poursuivre par l'étude des complexes chez les levures.

Les travaux alors réalisés ont démontré que cette technique d'identification des complexes pour un organisme est valide mais présente cependant quelques limites. La validité de la technique repose sur la mise en évidence de toutes les protéines impliquées dans un complexe, avec un poids moléculaire de complexe correct, tel le complexe de l'aspartate carbamyltransferase et le complexe de la succiny-CoA synthétase [Lasserre et al., 2006]. De plus, nous avons pu montrer que les spots alignés verticalement mais de forme différente ne sont pas forcément impliqué dans un même complexe. Pour augmenter la résolution des gels, le gradient du gel de polyacrylamide peut être modifié, de façon à voir la région considérée "grossie", par exemple en utilisant un gradient de 10-14% au lieu de 4-18%.

J'ai également validé la technique de l'électrophorèse BN/SDS chez *S. cerevisiae*. Par exemple, comme nous l'avons vu précédemment, chez *S. cerevisiae*, les deux sous-unités de l'hétérodimère Bmh1p/Bmh2p, respectivement de poids moléculaire de 30 kDa et 31 kDa, ont été identifiées en bonne position en seconde dimension (forme unitaire) et en première dimension (forme complexée) avec un poids moléculaire de ∼70 Kda (cf. fig. 6.1).

De même, chez *Y. lipolytica*, j'ai identifié à une position de ∼500 kDa, 11 des 14 protéines qui, par homologie avec les protéines de *S. cerevisiae*, se complexent pour former la partie catalytique 20S du protéasome 26S (cf. § 2.1.3.2) (cf. fig. 6.5). L'analyse de ces complexes est discutée au § 6.3.4.4.

Nous avons rencontré les mêmes limites concernant la stabilité du complexe. Par exemple, chez *Y. lipolytica*, trois des cinq sous-unités de la partie catalytique CF1 du complexe mitochondrial membranaire de l'ATP synthase F1F0 ont été mise en évidence comme complexées. Chez *S. cerevisiae*, L'ATP synthase F1F0 est constituée du core catalytique CF1 et du canal membranaire à proton CF0. CF1 se compose de 5 sous-unités avec la stœchiométrie suivante : $\alpha_3\beta_3\gamma\delta\epsilon$. Seules les sous-unités α, β et γ ont été observées à leur PM attendu. L'intensité du spot de γ était d'ailleurs plus faible que celle des deux premières (bien que la coloration argent ne soit pas fiable pour faire des analyses quantitatives, elle donne quand même quelques

indications). L'absence des sous-unités δ et ϵ pourrait s'expliquer soit par une instabilité dans la formation du complexe, soit par leur petite taille (respectivement 14,8 kDa et 6,7 kDa chez *S. cerevisiae*, cf. table 6.12) ce qui les situerait dans le bas ou en dehors du gel. La présence de ce complexe est discutée au § 6.3.4.4.

Par ailleurs, l'étape d'équilibration des protéines doit être réalisée avec soin car elle permet la séparation de chacun des composants des complexes. En effet, en seconde dimension, nous avons trouvé plusieurs homomultimères à plusieurs positions (cf. table 6.4), leur poids moléculaire étant des valeurs multiples de celui du monomère. Par exemple, chez *S. cerevisiae* (cf. fig.6.1), Act1p (41,7 kDa) a été trouvé aux positions de ∼40 kDa et ∼80 kDa, respectivement en tant que monomère et dimère. De même, Tsa1p (21,6 kDa) a été trouvé aux positions de ∼20 kDa et ∼40 kDa. Ces observations peuvent s'expliquer par une solubilisation de la bande du gel 1D pas assez efficace. Une autre hypothèse est l'enrichissement en complexes stables par cette technique. Dans ce cas, la solubilisation doit être beaucoup plus poussée pour dissocier les trop nombreux composants en interaction de façon stable.

Une autre limite concerne la mauvaise pénétration des complexes à très haut poids moléculaire dans le gel de première dimension. En seconde dimension, leurs sous-unités sont séparées mais mélangées. Ces complexes ne peuvent être identifiées par cette méthode car le gel d'électrophorèse doit contenir au minimum 3% de polyacrylamide pour être utilisable. Mais ils pourraient être caractérisés, par exemple, par la technique du TAP-tag (cf. § 2.4.4.2).

6.3.3 Identification des protéines

6.3.3.1 Analyses générales

À ce jour, j'ai identifié de façon sûre (au moins 2 scans/2 peptides) 50 protéines chez *S. cerevisiae* (cf annexe) localisées au moins une fois à une position correspondant à leur poids moléculaire attendue, excepté pour trois d'entre elles : Tfp1p (Ydl185wp), Tpi1p (Ydr050cp) et Wtm1p (Yor230wp). N'ayant pas identifié tous les spots en spectrométrie de masse et la position de ces trois protéines n'étant pas un multiple de leur poids moléculaire attendu, nous ne pouvons conclure s'il s'agit de produits de dégradation ou de modification post-traductionnelle.

Quant aux résultats pour *Y. lipolytica*, nous considérons ici l'ensemble des protéines identifiées quelles que soient les conditions de culture de *Y. lipolytica*. Nous avons ainsi identifié 114 protéines prédites *in silico* (au moins 2 scans/2 peptides) dont 96 protéines localisées au moins une fois à une position correspondant à leur poids moléculaire attendu.

Nous avons aussi identifié 4 protéines (cf. tab. 6.9 et fig. 6.3, en rouge) prédites comme n'ayant aucune similarité avec une séquence publiée dans les banques de donnée lors de l'annotation. Parmi ces quatre, YALI0A17020g1p (147,8 kDa, cf. fig. 6.5) est codée par un gène prédit avec un court intron. Étant localisé en seconde dimension à une position de ∼120 kDa, sachant qu'elle a migré en première dimension à une position de ∼500 kDa, cette protéine semble ainsi se complexer en homomultimère ou hétéromultimère (cf. § 6.3.4). La différence de poids moléculaire observée entre *in silico* et *in vivo* ne permet pas de dire si celui-ci est dû à une dégradation partielle de la protéine ou une mauvaise prédiction de la séquence codante

(notamment en ce qui concerne la justesse des motifs des sites d'épissage).

Même si ce nombre est faible (4 protéines sur 1029 protéines prédites "no similarity"), cela démontre que les critères de détection de gènes lors de la phase d'annotation sont justes.

Parmi les 114 protéines, 37 ont été prédites avec au moins un intron. Parmi elles, seules 5 protéines n'ont pas été observées à leur taille prédite. Ce point est repris au § 6.3.4.3.

Deux protéines ont été identifiées comme homomultimères d'après leur localisation en première dimension (cf. fig. 6.3) :

- YALI0F01672g1p : la localisation en seconde dimension correspond à celle de son poids moléculaire théorique (94,6 kDa), cette protéine migre dans un complexe de ∼800 kDa, ce qui corespondrait à un homo-octamère,
- YALI0B09647g1p : cette protéine de 62,9 kDa (poids moléculaire théorique et observé) migre, en première dimension, dans un complexe de ∼700 kDa, ce qui correspondrait à un complexe formé d'environ dix sous-unités.

La protéine YALI0B14641g1p (cf. fig. 6.3), de poids moléculaire théorique de 97,2 kDa, apparaît ici monomérique car elle est localisée à une position de ∼100kDa dans les deux dimensions.

Nous avons également identifié six protéines connues de *Y. lipolytica*, dont YALI0E05533g1p, une acétyltransferase intervenant lors de la première réaction de synthèse de la lysine et non présente chez *S. cerevisiae*.

6.3.3.2 Différence de milieux de culture

Nous pouvons comparer les gels obtenus avec des échantillons provenant des deux types de cultures cellulaires. Cette comparaison peut s'effectuer selon des critères de présence ainsi que l'intensité de coloration et la forme du spot.

Par exemple, nous avons observé une différence d'intensité pour la protéine YALI0F01672g1p. Pour une même quantité d'échantillon déposé, la protéine semble plus exprimée en présence d'huile (cf. fig. 6.3) que dans un milieu normal (cf. fig. 6.4). YALI0F01672g1p est annotée comme similaire à un précurseur de β-glucosidase de la levure *K. marxianus*. Elle appartient à la famille de protéines GLC.1656, définie par Génolevures, de profil phylétique "- -kdy" (pas d'équivalence chez les levures *S. cerevisiae* et *C. glabrata*) et de profil phylogénique "0 0 3 4 6" Cette famille regroupe les protéines similaires aux enzymes impliquées dans le métabolisme de l'acide cyanoaminé, du sucrose et de l'amidon ainsi que la synthèse du phénylpropanoïde (selon le KEGG, cette voie serait absente chez *S. cerevisiae*). Il apparaît ainsi que *Y. lipolytica* présente une expansion de cette fonction glucosidase alors que les levures Saccharomyces ont dû la perdre.

Nous observons également une intensité plus forte pour YALI0F16489g1p (cf. fig. 6.3 et fig. 6.4), une protéine similaire à une carboxypeptidase de *S. cerevisiae* impliquée dans le métabolisme de l'azote et localisée dans la vacuole, lieu de dégradation des protéines.

6.3.4 Identification de complexes protéiques

6.3.4.1 Implications dans plusieurs complexes

En seconde dimension, une même protéine peut apparaître à divers endroits alignés perpendiculairement au sens de migration 2D, et à la position de son poids moléculaire attendu. Cette protéine est impliquée dans différents complexes lorsque ses positions en 1D sont supérieures à son poids moléculaire.

Par exemple, la protéine YALI0C03443g1p (17,1 kDa) apparaît à quatre endroits (cf. fig6.3). Elle fait partie de la famille GLC.1967 (profils : phylétique "sck-y", phylogénique "2 1 2 0 3") qui contient une autre petite protéine chaperonne "heat shock protein". Ces protéines sont exprimées en réponse à des facteurs de stress tels que des variations de température. Elles s'assemblent et forment des tonneaux afin de protéger les protéines qui se dénaturent sous l'effet de chaleur et perdent ainsi leur fonctionnalité. YALI0C03443g1p (cf. fig 6.3) est localisée en première dimension à des positions de ∼70 kDa (homotetramère), ∼100 kDa, ∼170 kDa et ∼600 kDa. Pour ces trois derniers complexes, YALI0C03443g1p parait complexée avec diverses protéines (cf. tab. 6.6, complexes 3 et 4, fig. 6.7, complexe 5).

YALI0C06369g1p est également présente sous différentes formes. D'un poids moléculaire prédit de 35,8 kDa (poids moléculaire observé), cette glycéraldéhyde 3-phosphate déhydrogènase (GAPDH) a été localisé à cinq reprises (cf. fig. 6.3 et tab. 6.6 et 6.7) : ∼35 kDa (monomère), ∼100 kDa, ∼150 kDa, ∼350 kDa et ∼500 kDa.

Chez *S. cerevisiae*, nous remarquons également la présence de protéines impliquées dans divers complexes sous forme homomultimère ou hétéromultimère, tels que Eno2p, Tdh3p, Adh1p et Cdc19p.

6.3.4.2 Implications dans des complexes multimériques

Certaines protéines partageant la même voie métabolique, sont identifiées comme étant complexées. Par exemple, Eno2p et Tdh3p (ou Tdh2p) (l'analyse en spectrométrie de masse n'a pu déterminé si l'un ou l'autre ou les deux sont présents, leur séquence étant très proches et les motifs détectés étant identiques) sont deux enzymes de la glycolyse (cf. tab. 6.3, complexe 3).

Un autre complexe, localisé à ∼400 kDa, regroupe plusieurs enzymes de la glycolyse : Pgi1p, Tdh1p, Tdh2p ou Tdh3p, Ipp1p et Tpi1p (cf. tab. 6.3, complexe 2).

De même, Tal1p (37 kDa) et Tsa1p (21,5 kDa) (cf. tab. 6.3, complexe 10) appartiennent à un complexe de poids moléculaire ∼100 kDa en première dimension. Le fait d'identifier Tsa1p à ∼40 kDa nous permet d'en déduire que le complexe se compose de deux protéines Tsa1p et une protéine de Tal1p. Nous n'avons pas encore trouvé de point commun entre ces deux protéines si ce n'est leur localisation cytoplasmique.

De même, chez *Y. lipolytica*, plusieurs protéines sont impliquées dans divers complexes, telles YALI0C06369g1p et YALI0C03443g1p.

Nous pouvons également remarquer que les complexe 4, 5, 8 et 9 chez *Y. lipolytica*, sont constitués de protéines présentes dans la voie métabolique de la glycolyse (par similarité avec *S. cerevisiae*). Le complexe 6 de *Y. lipolytica* appartiendrait à la voie des pentoses phosphates.

Les résultats d'identification n'étant pas exhaustifs, ces résultats ne sont pas définitifs. Des analyses supplémentaires seront réalisées prochainement sur le protéome et le complexome de *Y. lipolytica*.

6.3.4.3 Validation *in vivo* des prédictions *in silico*

La difficulté d'analyse des données et le nombre insuffisant de complexes identifiés à ce jour ne nous permettent pas de comparer à grande échelle nos résultats expérimentaux avec ceux qui peuvent être par des méthodes bio-informatiques. Néanmoins, nous pouvons confirmer que la stratégie d'annotation adoptée par le consortium était bonne. Pour cela, plusieurs critères sont à considérer.

Le premier critère est la taille des protéines prédites. J'ai identifié 97 protéines, sur 114 protéines, localisées au poids moléculaire attendu.

Ce critère de taille implique nécessairement celui de la prédiction des introns, surtout chez *Y. lipolytica* qui a 14% de ses gènes avec au moins un intron. Ainsi, parmi ces 114 protéines, 37 sont prédites avec au moins un intron dont 5 localisées à une position non attendue.

Cette différence de poids moléculaire concerne aussi bien un poids observé inférieur de ∼10 kDa au poids attendu, qu'un poids observé supérieur. Plusieurs raisons peuvent expliquer cette différence : une dégradation partielle de la protéine, une mauvaise prédiction de la séquence codante, une modification post-traductionnelle, une structure 2D résiduelle, une charge électrique plus importante (ou moins importante) que la moyenne, une hydrophobicité différente de celle attendue en moyenne, un problème de solubilisation des protéines lors de l'étape d'équilibration...

Des analyses supplémentaires et complémentaires seraient nécessaires pour éliminer certaines possibilités, en l'occurrence la dégradation partielle et la solubilisation imparfaite des protéines.

YALI0F09185g1p, prédit avec un poids moléculaire de 51,7 kDa, est observé à une position de ∼60 kDa. Or des analyses expérimentales ([Strick et al., 1992, Strick et al., 1994], C. Neuvéglise, travaux en cours) ont mis en évidence un second intron en 5' du gène, donnant ainsi un poids moléculaire de 56 kDa, ce qui rejoint notre localisation observée de YALI0F09185g1p.

Le second critère est l'attribution de l'annotation fonctionnelle. En effet, certaines protéines complexées ensemble, possèdent une annotation commune pour une voie métabolique. Par exemple, les protéines Pgi1p, Tdh1p, Tdh2p ou Tdh3p, Ipp1p et Tpi1p ont chacune été prédites comme ayant une fonction présente dans la voie de la glycolyse et de la néoglucogénèse.

6.3.4.4 Comparaison de complexes entre levures

J'ai choisi de présenter deux complexes protéiques les mieux identifiés par électrophorèse BN/SDS chez *Y. lipolytica* et présentant des fonctions essentielles pour le cellule : le protéasome qui dégrade les protéines ; et le core catalytique de l'ATP synthase F1F0 qui synthétise la molécule d'ATP, source d'énergie pour la cellule.

Le protéasome En seconde dimension, nous détectons un complexe à haut poids moléculaire ($\geqslant$ 1200 kDa) pour lequel nous identifions 11 protéines. De par leur annotation, ces protéines sont toutes homologues (et souvent orthologues) aux protéines intervenant dans la composition du protéasome 26S de *S. cerevisiae* qui dégrade les protéines. Chez *S. cerevisiae*, le protéasome 26S se compose d'un sous-complexe catalytique 20S et de deux sous-complexes régulatoires 19S. La partie régulatoire se compose de 17 protéines dont 6 ATPases. Le sous-complexe catalytique 20S se compose de 4 anneaux heptamériques superposés de 14 sous-unités chacun : 7 sous-unités α (α1 à α7) et 7 sous-unités β (β1 à β7).

Comme nous l'avons vu au chapitre précédent, la présence de l'ensemble des éléments prédits constitutifs du protéasome 26S chez *Y. lipolytica* permet d'inférer la conservation fonctionnelle du protéasome 26S chez *Y. lipolytica*. En effet, le protéasome assure une fonction essentielle dans la cellule, il est soumis à une forte pression de sélection [Sorimachi et al., 1991]. L'étude des familles Génolevures auxquelles appartiennent les protéines du protéasome nous montre d'ailleurs que nombreuses sont celles qui n'ont qu'une seule protéine dans les cinq espèces Génolevures 2.

La table 6.11 présente les protéines de *S. cerevisiae* et *Y. lipolytica* constituant la partie catalytique 20S du protéasome, et identifiées et localisées à leur position adéquate (excepté pour une protéine) chez *Y. lipolytica*.

Le poids moléculaire du protéasome 20S (par homologie avec celui de *S. cerevisiae*) chez *Y. lipolytica* est ainsi prédit à $\sim$1523 kDa. Nous l'observons en effet à un haut poids moléculaire sur l'ensemble des gels (cf. fig 6.3 et 6.5).

L'ATP synthase F1F0 Comme nous l'avons vu au § 6.3.2, trois des cinq sous-unités de la partie catalytique CF1 de l'ATP synthase F1F0 ont été identifiées dans un complexe chez *Y. lipolytica*. L'ATP synthase F1F0 est un complexe membranaire interne de la mitochondrie qui synthétise l'ATP à partir du gradient de protons entretenus par la chaîne respiratoire.

Chez *S. cerevisiae*, le core catalytique CF1 se compose de 5 sous-unités avec la stœchiométrie suivante :$\alpha_3\beta_3\gamma\delta\epsilon$. La table 6.12 présente les protéines de *S. cerevisiae* et *Y. lipolytica* impliquées dans la partie catalytique CF1 de l'ATP synthase F1FO, identifiées et localisées à leur position adéquate chez *Y. lipolytica*.

Le poids moléculaire théorique de CF1 serait donc d'au moins 399,2 kDa (sans la sous-unité ϵ). D'après sa position sur le gel fig. 6.5, nous l'observons à une position de $\sim$400kDa. Le profil des familles Génolevures auxquelles appartiennent les protéines impliquées, nous montre la forte pression sélective imposée à ce complexe car, selon elles, il n'y qu'une séquence dans chacun des génomes comparable à celle de *S. cerevisiae* et pouvant assurer la même fonction. Ce qui nous conforte à rechercher activement la séquence manquante orthologue au YPL271W de *S. cerevisiae* chez *Y. lipolytica* qui, comme nous l'avons vu au chapitre précédent, est présente chez les trois autres espèces de levure *C. glabrata*, *K. lactis* et *D. hansenii*.

6.4 Conclusion et Perspectives

Les résultats obtenus par la technique de l'électrophorèse BN/SDS chez *Y. lipolytica* enrichit les données expérimentales pour des études de génomiques comparées chez les levures, autres que chez *S. cerevisiae*. Nous avons constaté tout d'abord que la prédiction de gènes sans similarité, de gènes avec introns, permet de valider l'annotation syntaxique. Ensuite, la mise en évidence de complexes multimériques tels que la sous-unité 20S du protéasome ou les complexes 8 et 9 de *Y. lipolytica* (membres annotés comme similaires à des protéines de la glycolyse chez *S. cerevisiae*, cf. tab. 6.7) permet dans ces cas de valider l'annotation fonctionnelle : les membres appartenant à un même complexe ont des annotations cohérentes entre elles. De même la mise en évidence d'interactions protéiques chez *Y. lipolytica*, prédites à partir de celles connues chez *S. cerevisiae* (*e.g.* pour la partie catalytique 20S), permet de valider ici l'annotation relationnelle. Ainsi, nous avons la confirmation que la stratégie adoptée pour l'annotation des levures lors de la phase Génolevures 2, bénéficie aux différents niveaux de l'annotation génomique.

Cette technique de gel d'électrophorèse en BN/SDS, qui ne requiert pas de manipulations moléculaires de la séquence génomique, nous permet ainsi d'obtenir à l'échelle de la cellule de nombreux résultats. La mise en évidence de nouvelles interactions protéine-protéine chez *S. cerevisiae* non détectées par les expériences de double hydride ou de purification en tandem montre la complémentarité de cette technique par rapport aux deux autres, couramment utilisées. Validée chez *E. coli* [Lasserre et al., 2006], *S. cerevisiae* et *Y. lipolytica*, elle peut être appliquée à d'autres organismes unicellulaires ou multicellulaires. Ce travail a fait l'objet de communications [Lasserre et al., 2005, Beyne et al., 2005].

Perspectives

Plusieurs perspectives s'ouvrent à partir du travail effectué.

Dans un premier temps, des analyses supplémentaires seraient nécessaires afin de lever les incertitudes sur la composition de certains complexes chez *Y. lipolytica*. Pour cela, les experiences en BN/SDS peuvent être poursuivies. D'autres techniques expérimentales peuvent être utilisées, en particulier le TAP-Tag (cf. § 2.4.4.2). Cette technique permet en effet d'isoler les différents partenaires d'interaction pour une protéine cible. Pour un complexe considéré, la protéine appât devra être choisie avec soin : privilégier une protéine impliquée dans ce seul complexe ou dans un minimum de complexes, afin de minimiser le nombre de faux positifs.

Ensuite, ces résultats devraient être croisés avec ceux obtenus par d'autres études de protéomique et complexomique qui seraient réalisées par d'autres équipes. Ils devraient également être comparés à d'autres analyses obtenues à partir de puces à ADN ou de banques d'ADNc, techniques utilisées par les membres du consortium Génolevures, telles que l'utilisation de puces à ADN ou de banques d'ADNc. L'utilisation de puce à ADN permettrait de s'assurer de l'existence de certains gènes non identifiés ici (*e.g.* les protéines non identifées de la partie catalytique 20S du protéasome). Le recours à une banque d'ADNc, telle que celle en cours de réalisation à AgroParisTech (centre de Grignon, Paris), permettrait de réajuster la taille des protéines. Cela lèverait l'incertitude sur la localisation des certaines protéines, et déter-

minerait alors si la variation par rapport à la taille attendue est due à une dégradation, à des modifications post-traductionnelles ou à une erreur de prédiction de séquence.

Les variations d'expression protéomique en fonction de la source carbonnée (présence d'huile), pourraient être étudiées par la méthode du iTRAQ (isobaric Tagging Reagent) [Ross et al., 2004]. Cette méthode de protéomique quantitative permet en effet de connaître simultanément les quantités d'un mélange complexe protéique au cours de quatre expériences. Dans notre cas, si les conditions de milieux de croissance constituaient ainsi deux expériences, nous pourrions en plus faire une étude de cinétique en étudiant, pour chacune des conditions, les quantités protéiques en fonction de la croissance ou de l'activité lysosomale. Nous pourrions modéliser ainsi les changements métabolomiques en fonction des sources nutritives et de la croissance des levures.

Cette étude métabolomique pourrait être couplée à une étude transcriptomique en analysant l'expression des gènes par l'utilisation de puce à ADN.

À plus long terme, la mise en évidence de complexes protéiques, par méthodes *in silico* et *in vivo* vues précédemment (*e.g.* gel d'électrophorèse en BN et SDS, iTRAQ, TAP-Tag...), pourraient être étudiées chez les autres levures du projet Génolevures, afin de préciser l'évolution de ces complexes et leur implication dans les voies métaboliques.

TAB. 6.3 – Complexes multimériques chez *S. cerevisiae*.

Toutes les protéines ont été identifiées avec au moins 2 peptides différents sauf indication contraire. Les IPP présentes dans les banques de données Intact et BioGrid sont précisées (num. cplx : numéro du complexe, % couv : pourcentage de couverture des peptides identifiés par rapport à la longueur totale de la protéine, PM th. : poids moléculaire théorique (kDa)).

num. cplx.	% couv.	gène	protéine	PM th.	fonction protéique	IPP connue	remarques
1	20	YAL038W	Cdc19	54,5	Pyruvate kinase 1		
1	24	YGR254W	Eno1p	46,8	Enolase 1		
2	34	YBR011C	Ipp1p	32,3	Pyrophosphatase		
2	29	YBR196C	Pgi1p	61,3	Glucose-6-phosphate isomerase		
2	35	YDR050C	Tpi1p	26,8	Triose phosphate isomerase		PM incorrect
2	21	YGR192C	Tdh3p	35,7	Glyceraldehyde-3-phosphate dehydrogenase		
2	8	YJL052W	Tdh1p	35,7	Glyceraldehyde-3-phosphate dehydrogenase	Tdh2p	
2	11	YJR009C	Tdh2p	35,8	Glyceraldehyde-3-phosphate dehydrogenase	Tdh1p	
3	21	YGR192C	Tdh3p	35,7	Glyceraldehyde-3-phosphate dehydrogenase 3		
3	11	YJR009C	Tdh2p	35,8	Glyceraldehyde-3-phosphate dehydrogenase 2		
3	32	YHR174W	Eno2p	46,9	Enolase 2		
4	21	YDR385W	Eft2p	93,3	Translation elongation factor 2	Eft1p	
4	21	YOR133W	Eft1p	93,3	Translation elongation factor 2	Eft2p	
5	9	YMR303C	Adh2p	36,7	Alcohol dehydrogenase 2	Adh1p	
5	23	YOL086C	Adh1p	36,8	Alcohol dehydrogenase 1	Adh2p	
6	18	YMR116C	Asc1p	34,8	Guanine nucleotide-binding protein subunit β-like protein		
6	5	YGR155W	Cys4	56	Cystathionine beta-synthase	Tef1p	
6	8	YPR080W	Tef1p	50	Translation Elongation factor 1-alpha	Cys4	
6	8	YBR118W	Tef2p	50	Translational elongation factor 1 alpha		
7	12	YDR099W	Bmh2p	31	signalisation cellulaire	Bmh1p	
7	9	YER177W	Bmh1p	30	signalisation cellulaire	Bmh2p	
8	7	YJL138C	Tif2p	44,7	ATP-dependent RNA helicase eIF4A	Tif1p	
8	7	YKR059W	Tif1p	44,7	ATP-dependent RNA helicase eIF4A	Tif2p	
9	25	YLL024C	Ssa2p	69,5	Heat shock protein SSA2		
9	6	YNL209W	Ssb2p	66,6	Heat shock protein SSB2		
10	7	YLR354C	Tal1p	37	Transaldolase		
10	25	YML028W	Tsa1p	21,5	Peroxidase	observée à 40 kDa	

TAB. 6.4 – Complexes homodimériques chez *S. cerevisiae*.

Toutes les protéines ont été identifiées avec au moins 2 peptides différents. Les IPP présentes chez Intact et BioGrid sont précisées (num. cplx : numéro du complexe, % couv : pourcentage de couverture des peptides identifiés par rapport à la longueur totale de la protéine, PM th. : poids moléculaire théorique (kDa)).

num. cplx.	% couv.	gène	protéine	PM th.	fonction protéique	IPP connue	remarques
10	10	YFL039C	Act1p	41,7	Actine		homodimere
11	16	YLR109W	Ahp1p	19,1	Peroxiredoxin type-2	homodimère	homodimere
12	25	YML028W	Tsa1p	21,6	Peroxiredoxin TSA1	homodimère	homodimère

TAB. 6.5 – Protéines identifiées chez *S. cerevisiae*.
Toutes les protéines ont été identifiées avec au moins 2 peptides différents (num. cplx : numéro du complexe, % couv : pourcentage de couverture des peptides identifiés par rapport à la longueur totale de la protéine, PM th. : poids moléculaire théorique (kDa)).

num. cplx.	% couv.	gène	protéine	PM th.	fonction protéique	remarques
13	12	YDR099W	Bmh2p	31	signlisation cellulaire	
14	21	YGR192C	Tdh3p	35,7	Glyceraldehyde-3-phosphate dehydrogenase 3	
15	11	YJR009C	Tdh2p	35,8	Glyceraldehyde-3-phosphate dehydrogenase 2	
16	20	YAL038W	Cdc19	54,5	Pyruvate kinase 1	
17	8	YBR118W	Tef2p	50	Translational elongation factor 1 alpha	
18	25	YLR044C	Pdc1p	61,5	Pyruvate decarboxylase isozyme 1	
19	8	YPR080W	Tef1p	50	Translation Elongation factor 1-alpha	
20	8	YJR121W	Atp2p	54,8	Beta subunit of the F1 sector of mitochondrial F1F0 ATP synthase	
21	18	YAL005C	Ssa1p	69,7	Heat shock protein Ssa1p	
22	12	YAL012W	Cys3p	42,5	Cystathionine gamma-lyase	
23	44	YCR012W	Pgk1p	44,7	Phosphoglycerate kinase	
24	5	YDL185W	Tfp1p	118,6	Vacuolar ATP synthase catalytic subunit A	PM incorrect
25	5	YDL229W	Ssb1p	66,6	Heat shock protein SSB1	
26	7	YGL202W	Aro8p	56,2	Aromatic amino acid aminotransferase 1	
27	13	YGR234W	Yhb1p	44,6	Flavohemoprotein, Nitric oxide oxidoreductase	
28	10	YJR016C	Ilv3p	62,9	Dihydroxy-acid dehydratase, mitochondrial	
29	16	YKL080W	Vma5p	44,2	Vacuolar ATP synthase subunit C	
30	1.5	YKL157W	Ape2p	105,6	Aminopeptidase 2, mitochondrial precursor	1 peptide
31	5	YLR058C	Shm2p	52,2	Serine hydroxymethyltransferase	
32	28	YLR259C	Hsp60p	60,7	Heat shock protein 60, mitochondrial precursor	
33	5	YLR303W	Met17p	48,7	Methionine and cysteine synthase	
34	7	YLR354C	Tal1p	37	Transaldolase	
35	13	YMR120C	Ade17p	65,3	Bifunctional purine biosynthesis protein	
36	23	YMR186W	Hsc82p	80,9	ATP-dependent molecular chaperone	
37	6	YOR027W	Sti1p	66,3	Heat shock protein STI1	
38	16	YOR230W	Wtm1p	48,4	Transcriptional modulator	PM incorrect
39	25	YOR375C	Gdh1p	49,6	NADP-specific glutamate dehydrogenase 1	
40	4	YPL240C	Hsc82p	81,4	ATP-dependent molecular chaperone HSP82	
41	20	YPR074C	Tkl1p	73,8	Transketolase 1	
42	14	YKL060C	Fba1p	39,6	Fructose 1,6-bisphosphate aldolase	
43	29	YKL152C	Gpm1p	27,6	Phosphoglycerate mutase	

TAB. 6.6 – Complexes multimériques chez *Y. lipolytica*.
Toutes les protéines ont été identifiées avec au moins 2 peptides différents sauf indication contraire. Les IPP présentes dans les banques de données Intact et BioGrid sont précisées (cplx : numéro du complexe, couv : pourcentage de couverture des peptides identifiés par rapport à la longueur totale de la protéine, PMth : poids moléculaire théorique (kDa), PMo : poids moléculaire observé (kDa), I : nombre d'intron(s)).

cplx	protéine	couv	gène	PMt	PMo	fonction protéique	IPP connue	I	remarques
1	YALI0C06039g1p	9	YALI0C06039g	26.9	27	YGL011c SCL1 20S proteasome subunit YC7ALPHA/Y8	protéasome 20S	1	
1	YALI0F06314g1p	16	YALI0F06314g	27.3	27	YML092c PRE8 20S proteasome subunit Y7 (alpha2)	protéasome 20S	1	
1	YALI0C19382g1p	20	YALI0C19382g	27.8	27	YGR135w PRE9 20S proteasome subunit Y13 (alpha3) P7.1.f7.1	protéasome 20S		
1	YALI0B14267g1p	10	YALI0B14267g	27.3	28	YOL038w PRE6 20S proteasome subunit (alpha4) P7.1.f7.1	protéasome 20S	1	
1	YALI0E02794g1p	16	YALI0E02794g	27.8	30	YGR253c PUP2 20S proteasome subunit(alpha5)	protéasome 20S	1	
1	YALI0C17325g1p	41	YALI0C17325g	31.1	30	YMR314w PRE5 20S proteasome subunit(alpha6)	protéasome 20S	1	
1	YALI0F07469g1p	8.3	YALI0F07469g	27.5	27	YOR362c PRE10 20S proteasome subunit C1 (alpha7)	protéasome 20S	1	
1	YALI0D14058g1p	7	YALI0D14058g	27.9	15	YOR157c PUP1 20S proteasome subunit (beta2)	protéasome 20S		PM incorrect
1	YALI0B11374g1p	11	YALI0B11374g	24.3	27	YER012w PRE1 20S proteasome subunit C11(beta4) P4.15.f3.1	protéasome 20S		
1	YALI0B15224g1p	3	YALI0B15224g	31.1	25	YPR103W PRE2 Proteasome component precursor	protéasome 20S		1 peptide
1	YALI0D06523g1p	8	YALI0D06523g	27.7	25	YBL041w PRE7 20S proteasome subunit(beta6)	protéasome 20S	1	
2	YALI0F03179g1p	17	YALI0F03179g	58.1	50	YBL099w ATP1 F1F0-ATPase complex F1 alpha subunit	CF1 de l'ATP synthase F1FO	1	PM incorrect
2	YALI0B03982g1p	34	YALI0B03982g	59.7	50	YJR121W ATP synthase beta chain mitochondrial precursor	CF1 de l'ATP synthase F1FO	1	
2	YALI0F02893g1p	35	YALI0F02893g	31	30	YBR039w ATP3 F1F0-ATPase complex F1 gamma subunit	CF1 de l'ATP synthase F1FO	1	
3	YALI0C03443g1p	15	YALI0C03443g	17.1	17	tr|Q89C37 B. japonicum Blr7961 protein			homomultimère
3	YALI0C16621g1p	6	YALI0C16621g	23.1	23	YHR008C SOD2 superoxide dismutase			1 peptide
4	YALI0C03443g1p	15	YALI0C03443g	17.1	17	tr|Q89C37 B. japonicum Blr7961 protein			homomultimère
4	YALI0C06369g1p	21	YALI0C06369g	35.8	35	tr|Q9Y796 C. curvatus Glyceraldehyde 3-phosphate dehydrogenase			
4	YALI0B05346g1p	6	YALI0B05346g	23.3	23	DEHA0E22990g	1 peptide		

TAB. 6.7 – Complexes multimériques chez *Y. lipolytica* (suite).

cplx	protéine	couv	gène	PMt	PMo	fonction protéique	IPP connue	I	remarques
5	YALI0C03443g1p	15	YALI0C03443g	17.1	17	tr|Q89C37 B. japonicum Blr7961 protein			homomultimère
5	YALI0C06369g1p	21	YALI0C06369g	35.8	35	tr|Q9Y796 C. curvatus Glyceraldehyde 3-phosphate dehydrogenase			
6	YALI0C06369g1p	21	YALI0C06369g	35.8	35	tr|Q9Y796 C. curvatus Glyceraldehyde 3-phosphate dehydrogenase			glycolyse
6	YALI0E06479g1p	33.1	YALI0E06479g	76.2	75	YPR074c TKL1 transketolase 1 or YBR117c TKL2 transketolase 2		1	
7	YALI0E14190g1p	35	YALI0E14190g	34.6	35	YKL085w MDH1 malate dehydrogenase		1	cycle TCA
7	YALI0B06413g1p	21	YALI0B06413g	46.2	45	YHR179w OYE2 NADPH dehydrogenase			
8	YALI0D12400g1p	26	YALI0D12400g	45.8	45	Phosphoglycerate kinase (identifiée)		1	glycolyse
8	YALI0F18590g1p	8	YALI0F18590g	36.2	36	YOR120w GCY1 aldo/keto reductase or YDR368w YPR1 aldo/keto reductase			glycolyse
9	YALI0E26004g1p	28	YALI0E26004g	39.9	40	YKL060c FBA1 fructose-bisphosphate aldolase		1	glycolyse
9	YALI0C06369g1p	21	YALI0C06369g	35.8	35	tr|Q9Y796 C. curvatus Glyceraldehyde 3-phosphate dehydrogenase			glycolyse
9	YALI0B02728g1p	39	YALI0B02728g	27.5	27	YKL152C Phosphoglycerate mutase 1			glycolyse

TAB. 6.8 – Complexes homomultimériques chez *Y. lipolytica*.

Toutes les protéines ont été identifiées avec au moins 2 peptides différents sauf indication contraire. Les IPP présentes dans les banques de données Intact et BioGrid sont précisées (cplx : numéro du complexe, couv : pourcentage de couverture des peptides identifiés par rapport à la longueur totale de la protéine, PMt : poids moléculaire théorique (kDa), PMo : poids moléculaire observé (kDa), I : nombre d'intron).

cplx	protéine	couv	gène	PMt	PMo	fonction protéique	I	remarques
10	YALI0F30613g1p	23	YALI0F30613g	33.3	35	tr|Q9P5M9 S. pombe putative class II aldolase		homotrimère
11	YALI0A17020g1p	9	YALI0A17020g	147.8	140	no similarity	1	multimère
12	YALI0F01672g1p	18	YALI0F01672g	94.5	95	sp|P07337 K. marxianus Beta-glucosidase precursor		homo-octamère
13	YALI0B09647g1p	34	YALI0B09647g	62.9	60	YHR037W delta-1-pyrroline-5-carboxylate dehydrogenase		homodecamère ?
14	YALI0C03443g1p	15	YALI0C03443g	17.1	17	tr|Q89C37 B. japonicum Blr7961 protein		homomultimère, hétéomultimère
15	YALI0F16819g1p	45	YALI0F16819g	47.3	50	YHR174w ENO2 enolase or YGR254w ENO1 enolase	1	homomultimère
16	YALI0F09185g1p	10	YALI0F09185g	51.7	60	Pyruvate kinase (identifié)	1	homotétramère

TAB. 6.9 – Protéines identifiées chez *Y. lipolytica*.

protéine	couv	gène	PMt	PMo	fonction protéique	I	remarques
YALI0A00132g1p	7	YALI0A00132g	66.1	30	YNL209w SSB2 heat shock protein		PM incorrect
YALI0A00352g1p	4	YALI0A00352g	93.3	100	YOR133w EFT1 translation elongation factor eEF2	1	
YALI0A00847g1p	7	YALI0A00847g	39.4	35	tr|O59826 S. pombe Putative potassium channel subunit		
YALI0A14806g1p	5	YALI0A14806g	52.8	50	sp|P11913 N. crassa Mitochondrial processing peptidase beta subunit		
YALI0A15950g1p	15	YALI0A15950g	91.7	100	YCL030c HIS4 phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase/histidinol dehydrogenase		
YALI0A19074g1p	12	YALI0A19074g	168	170	YPR184W GDB1 glycogen debranching enzyme	1	
YALI0A21439g1p	5	YALI0A21439g	43.2	45	wi|NCU09800.1 N. crassa NCU09800.1 putative dioxygenase		
YALI0B00924g1p	19	YALI0B00924g	17.8	17	no similarity		
YALI0B03124g1p	11	YALI0B03124g	29.1	30	tr|Q8NKC1 S. pombe Putative O-methyltransferase		
YALI0B03366g1p	4	YALI0B03366g	21.1	20	YLR178C DKA1 Carboxypeptidase Y inhibitor	1	2 pep ? 1 pep ?
YALI0B04312g1p	17	YALI0B04312g	58.1	60	tr|O74247 P. pastoris and YJL153c INO1 myo-inositol-1-phosphate synthase		
YALI0B05522g1p	10	YALI0B05522g	52.9	53	tr|Q96TK5 C. immitis Aspartyl aminopeptidase		
YALI0B08184g1p	16	YALI0B08184g	64	65	tr|Q8NJN1 A. niger and YJR010w MET3 ATP sulfurylase		
YALI0B08536g1p	31	YALI0B08536g	40.4	45	sp|Q04677 C. tropicalis Acetyl-CoA acetyltransferase IB) Peroxisomal	1	
YALI0B08921g1p	11	YALI0B08921g	25.5	25	sp|O13401 C. albicans Superoxide dismutase [Mn] mitochondrial precursor		
YALI0B10406g1p	12	YALI0B10406g	37.2	27	tr|Q9V6U5 D. melanogaster LD24265p	1	PM incorrect
YALI0B13970g1p	2	YALI0B13970g	67.5	60	YIL155c GUT2 glycerol-3-phosphate dehydrogenase mitochondrial		1 peptide
YALI0B14377g1p	7	YALI0B14377g	31.1	35	14-3-3 protein promoting filamentous growth (identifiée)	1	
YALI0B14641g1p	14	YALI0B14641g	97.2	97	YKL157w APE2 aminopeptidase yscII	1	monomère
YALI0B15598g1p	17	YALI0B15598g	53.5	53	YHR183w GND1 6-phosphogluconate dehydrogenase P2.360.f2.1 or YGR256w GND2 phosphogluconate dehydrogenase	1	
YALI0B16104g1p	2	YALI0B16104g	58.3	62	YMR217w GUA1 GMP synthase (glutamine-hydrolyzing)		1 peptide
YALI0B19580g1p	19	YALI0B19580g	74.9	75	YOL057W DPP3 Probable dipeptidyl-peptidase III		monomère
YALI0B21846g1p	4	YALI0B21846g	44.2	27	sp|Q00415 C. immitis 4-hydroxyphenylpyruvate dioxygenase		PM incorrect
YALI0B21868g1p	12	YALI0B21868g	25	25	tr|Q20683 C. elegans F52H3.5 protein		
YALI0B23100g1p	7	YALI0B23100g	34.5	35	DEHA0E01430g		
YALI0B23430g1p	18	YALI0B23430g	47.4	45	sp|P31225 C. albicans Corticosteroid-binding protein		
YALI0C01221g1p	29	YALI0C01221g	24.4	30	tr|Q9P3P6 N. crassa Related to 26s proteasome subunit p28		
YALI0C01859g1p	12	YALI0C01859g	56.6	55	sp|Q02253 R. norvegicus Methylmalonate-semialdehyde dehydrogenase mitochondrial precursor		
YALI0C04433g1p	14	YALI0C04433g	46.7	120	sp|Q92413 E. nidulans Ornithine aminotransferase		PM incorrect (trimere)
YALI0C06171g1p	9	YALI0C06171g	36.6	40	tr|Q884Q9 P. syringae Oxidoreductase zinc-binding		
YALI0C06776g1p	8	YALI0C06776g	52.9	53	sp|P55250 R. oryzae Fumarate hydratase mitochondrial precursor		

protéine	couv	gène	PMt	PMo	fonction protéique	I	remarques
YALI0C07953g1p	11	YALI0C07953g	80.3	85	sp|P46598 C. albicans CaHSC82 Heat shock protein 90 homolog		
YALI0C09141g1p	10	YALI0C09141g	50.1	30	Elongation factor 1-alpha (identifiée)	1	dimère observé
YALI0C10494g1p	22.1	YALI0C10494g	54.4	50	YKL103C LAP4 Vacuolar aminopeptidase I precursor		PM incorrect
YALI0C11385g1p	21	YALI0C11385g	11.4	30	YNL030w HHF2 histone H4	1	PM incorrect
YALI0C16797g1p	9	YALI0C16797g	58.4	50	tr|Q96VP9 G. intraradices Probable acyl-CoA dehydrogenase	1	PM incorrect
YALI0C16885g1p	17	YALI0C16885g	60	60	Isocitrate lyase (identifiée)		
YALI0C16973g1p	3	YALI0C16973g	44	45	DEHA0G11330g and YDR346c		1 peptide
YALI0C17831g1p	5	YALI0C17831g	54.7	55	YOL049w GSH2 Glutathione synthetase		
YALI0C20449g1p	8	YALI0C20449g	27.7	27	YGR207c ETF-BETA electron-transferring flavoprotein beta chain		
YALI0C21362g1p	13	YALI0C21362g	41.4	30	YGR234w YHB1 flavohemoglobin		PM incorrect
YALI0C23408g1p	2	YALI0C23408g	60.4	60	YJR016c ILV3 Dihydroxy-acid dehydratase mitochondrial precursor		1 peptide
YALI0C24101g1p	8	YALI0C24101g	130.4	110	YGL062w PYC1 pyruvate carboxylase 1	1	
YALI0C24255g1p	17	YALI0C24255g	32.1	35	YMR096w SNZ1 stationary phase protein		
YALI0D04268g1p	13	YALI0D04268g	32.4	30	YKR066c CCP1 cytochrome-c peroxidase precursor		
YALI0D08184g1p	6	YALI0D08184g	70.3	35	YER103w SSA4 heat shock		PM incorrect
YALI0D08272g1p	8	YALI0D08272g	42.2	42	Actine (identifiée)	1	
YALI0D12386g1p	12	YALI0D12386g	39.4	40	YGL157w and DEHA0A06347g		
YALI0D16753g1p	35	YALI0D16753g	35.9	35	tr|Q8TG27 T. emersonii Malate dehydrogenase precursor mitochondrial		
YALI0D21530g1p	4	YALI0D21530g	56.6	60	CA1245|CaIMH3 Candida albicans IMP dehydrogenase	1	
YALI0D22352g1p	6	YALI0D22352g	70.1	35	YER103w SSA4 heat shock protein		PM incorrect, degradation
YALI0D26367g1p	4	YALI0D26367g	52	52	YHR018c ARG4 arginosuccinate lyase		
YALI0D27126g1p	12	YALI0D27126g	34.2	35	YDR353W TRR1 Thioredoxin reductase 1		
YALI0E00484g1p	20	YALI0E00484g	93.1	110	no similarity		
YALI0E02684g1p	7	YALI0E02684g	51.7	42	YNR001c CIT1 citrate (si)-synthase mitochondrial possible transmembrane segment	1	PM incorrect
YALI0E05467g1p	4	YALI0E05467g	24.3	25	tr|Q88LQ9 P. putida PP1870 Thiopurine s-methyltransferase		1 peptide
YALI0E05511g1p	14	YALI0E05511g	24.5	17	no similarity		PM incorrect
YALI0E05533g1p	10	YALI0E05533g	43.9	42	Lysine acetyltransferase (identifié)		
YALI0E07073g1p	14	YALI0E07073g	31.7	35	YFR047c putative nicotinate-nucleotide pyrophosphorylase [carboxylating]		
YALI0E12463g1p	12.9	YALI0E12463g	38.1	40	tr|O74230 Candida sp.HA167 XDH Xylitol dehydrogenase		
YALI0E12595g1p	21	YALI0E12595g	61.7	60	tr|Q9V9A7 D. melanogaster Putative propionyl-CoA carboxylase beta chain mitochondrial precursor		
YALI0E13420g1p	10	YALI0E13420g	34.6	35	tr|Q9HZP7 P. aeruginosa Electron transfer flavoprotein alpha-subunit		
YALI0E13464g1p	3.7	YALI0E13464g	51.9	50	DEHA0F26642g and YFR006w		1 peptide

protéine	couv	gène	PMt	PMo	fonction protéique	I	remarques
YALI0E16753g1p	11	YALI0E16753g	23.2	25	YJR133w XPT1 xanthine phosphoribosyl transferase		
YALI0E16929g1p	9	YALI0E16929g	48.8	50	YDR148c KGD2 2-oxoglutarate dehydrogenase complex E2 component and tr|Q9UWE0 A. fumigatus Dihydrolipoamide succinyltransferase		
YALI0E17787g1p	35	YALI0E17787g	37.3	37	Alcohol dehydrogenase 2 (identifié)		
YALI0E18238g1p	10	YALI0E18238g	54.8	52	YGR019w UGA1 4-aminobutyrate aminotransferase (GABA transaminase)		
YALI0E19184g1p	10	YALI0E19184g	72.3	72	tr|Q8RY11 A. thaliana AT3g05350/T12H1_32 putative X-pro dipeptidase		
YALI0E20977g1p	19	YALI0E20977g	55.9	56	YGL202w ARO8 aromatic amino acid aminotransferase I		
YALI0E25025g1p	5	YALI0E25025g	21.7	20	YBR084ca RPL19B 60S large subunit ribosomal protein L19		1 peptide
YALI0E27962g1p	19	YALI0E27962g	99.1	100	YLL026w HSP104 heat shock protein		
YALI0E29975g1p	15	YALI0E29975g	51.8	35	YDR044w HEM13 coproporphyrinogen III oxidase		PM incorrect
YALI0E31911g1p	16	YALI0E31911g	24.3	30	YPL220w SSM1A ribosomal protein		
YALI0E32901g1p	2	YALI0E32901g	57.9	20	YKL217W JEN1 Carboxylic acid transporter protein homolog		1 peptide
YALI0E34265g1p	5	YALI0E34265g	60.4	60	YGR088w CTT1 catalase T cytosolic		
YALI0E34749g1p	14	YALI0E34749g	59.9	60	YGR088w CTT1 catalase T cytosolic		
YALI0E34793g1p	4	YALI0E34793g	71.4	70	tr|Q8X097 N. crassa Probable ATP citrate lyase subunit 1		
YALI0E35046g1p	3	YALI0E35046g	74.2	33	YER103w SSA4 heat shock protein		PM incorrect
YALI0F00682g1p	46	YALI0F00682g	26.5	27	YDR533c hypothetical protein		
YALI0F02805g1p	39	YALI0F02805g	60.5	60	YLR259c HSP60 heat shock protein - chaperone mitochondrial		
YALI0F05214g1p	36	YALI0F05214g	26.8	27	YDR050c TPI1 triose-phosphate isomerase	1	
YALI0F07711g1p	31	YALI0F07711g	61.9	60	YBR196c PGI1 glucose-6-phosphate isomerase		
YALI0F09229g1p	33	YALI0F09229g	17.1	15	YKL067w YNK1 nucleoside diphosphate kinase	1	
YALI0F09669g1p	15	YALI0F09669g	23.9	35	YAL003w EFB1 translation elongation factor eEF1beta	1	PM incorrect, haut
YALI0F15587g1p	28	YALI0F15587g	35.7	35	YLR354c TAL1 transaldolase	1	
YALI0F16489g1p	18	YALI0F16489g	64.2	65	YJL172w CPS1 Gly-X carboxypeptidase YSCS precursor		
YALI0F17842g1p	19	YALI0F17842g	52.5	53	YFR044c Glutamate carboxypeptidase-like protein	1	
YALI0F17974g1p	23	YALI0F17974g	52.7	55	YNL239w Cysteine proteinase 1		
YALI0F20592g1p	2	YALI0F20592g	51.8	50	YKL103c LAP4 aminopeptidase yscI precursor vacuolar		1 peptide
YALI0F20790g1p	16	YALI0F20790g	37.5	37	YGL040c HEM2 porphobilinogen synthase		
YALI0F25289g1p	6	YALI0F25289g	70.4	35	YER103w SSA4 heat shock protein		PM incorrect
YALI0F25883g1p	21	YALI0F25883g	11.4	30	YNL030w HHF2 histone H4		PM incorrect, dimere
YALI0F26191g1p	27	YALI0F26191g	54.9	55	YBR006w UGA2 succinate semialdehyde dehydrogenase		
YALI0F27049g1p	36	YALI0F27049g	21.1	21	YIL138c TPM2 tropomyosin		
YALI0F31999g1p	14	YALI0F31999g	63	63	YPR006c ICL2 non-functional isocitrate lyase		

TAB. 6.10 – Observation d'une variation de poids moléculaire pour 5 protéines de *Y. lipolytica*

protéine	PM théorique (kDa)	PM observé (kDa)	nbr intron
YALI0B10406g1p	37,2	25	1
YALI0C11385g1p	11,4	30	1
YALI0C16797g1p	58,4	50	1
YALI0C24101g1p	130,4	110	1
YALI0F09185g1p	51,7	60	2

TAB. 6.11 – Composition de la partie catalytique 20S du protéasome 26S chez *S. cerevisiae* et *Y. lipolytica*.

Le profil phylétique et celui phylogénétique sont indiqués pour chaque famille Génolevures. Seuls les protéines $\beta 2$ de *S. cerevisiae* et *Y. lipolytica* appartiennent à des familles Génolevures distinctes. (ssu : sous-unité protéique, Sc : *S. cerevisiae*, Yl : *Y. lipolytica*. PM th : poids moléculaire théorique (kDa)).

ssu	gène Sc	protéine Sc	PM th	protéine Yl	PM th	famille	remarque
$\alpha 1$	YGL011C	Scl1p	28,0	YALI0C06039g1p	26,9	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 2$	YML092C	Pre8p	27,1	YALI0F06314g1p	27,2	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 3$	YGR135W	Pre9p	28,7	YALI0C19382g1p	27,8	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 4$	YOL038W	Pre6p	28,4	YALI0B14267g1p	27,3	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 5$	YGR253C	Pup2p	28,6	YALI0E02794g1p	27,8	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 6$	YMR314W	Pre5p	25,6	YALI0C17325g1p	31,1	GLS.89 sckdy (7 7 7 7 7)	
$\alpha 7$	YOR362C	Pre10p	31,5	YALI0F07469g1p	27,6	GLS.89 sckdy (7 7 7 7 7)	
$\beta 1$	YJL001W	Pre3p	23,5	YALI0F14861g1p	23,6	GLS.79 sckdy (2 2 2 2 2)	non identifiée
$\beta 2$	YOR157C	Pup1p	28,2	YALI0D14058g1p	27,9	GLS.79 sckdy (2 2 2 2 2)	position basse
$\beta 3$	YER094C	Pup3p	22,6	YALI0F22341g1p	22,8	GLS.96 sckdy (1 1 1 1 1)	non identifiée
$\beta 4$	YER012W	Pre1p	22,5	YALI0B11374g1p	24,3	GLS.97 sckdy (1 1 1 1 1)	
$\beta 5$	YPR103W	Pre2p	31,6	YALI0B15224g1p	31,1	GLS.98 sckdy (1 1 1 1 1)	1 peptide
$\beta 6$	YBL041W	Pre7p	26,9	YALI0D06523g1p	26,8	GLS.99 sckdy (1 1 1 1 1)	
$\beta 7$	YFR050C	Pre4p	29,4	YALI0E32505g1p	29,7	GLS.100 sckdy (1 1 1 1 1)	non identifiée

TAB. 6.12 – Composition de la partie catalytique CF1 de l'ATP synthase F1F0 chez *S. cerevisiae* et *Y. lipolytica*.

Le profil phylétique et celui phylogénétique sont indiqués pour chaque famille Génolevures. Seule la sous-unité ϵ de *S. cerevisiae* n'a pas d'homologue chez *Y. lipolytica*. (ssu : sous-unité protéique, Sc : *S. cerevisiae*, Yl : *Y. lipolytica*. PM th : poids moléculaire théorique (kDa)).

ssu	gène Sc	protéine Sc	PM th	protéine Yl	PM th	famille	identifiée
α	YBL099W	Atp1p	58,6	YALI0F03179g1p	58,1	GLR.2726 sckdy (1 1 1 1 1)	oui
β	YJR121W	Atp2p	54,8	YALI0B03982g1p	59,7	GLR.2609 sckdy (1 1 1 1 1)	oui
γ	YBR039W	Atp3p	34,5	YALI0F02893g1p	31	GLR.712 sckdy (1 1 1 1 1)	oui
δ	YDL004W	Atp16p	17	YALI0D22022g1p	14,8	GLR.945 sckdy (1 1 1 1 1)	-
ϵ	YPL271W	Atp15p	6,7	?		GLC.2073 sckd- (1 1 1 1 0)	-

Chapitre 7

Conclusion générale

À Notre connaissance, les stratégies d'annotation d'un organisme ne s'assurent pas que cette dernière décrit un système fonctionnel biologique cohérent, même si elle comporte très peu d'erreurs grâce à la participation d'experts biologiques. Or, la qualité de l'annotation se définit par la complétude (identification de tous les éléments d'intérêt d'un génome) et l'absence d'erreur (sens correct des éléments identifiés).

Nous avons ainsi proposé une démarche pour vérifier et accroître la qualité de l'annotation d'un génome selon ces deux critères. Cette démarche avait pour cadre le projet Génolevures, projet de génomique comparée à grande échelle chez les levures Hémiascomycètes. Elle procède par la définition d'un ensemble de règles de logique pour la vérification de la cohérence de l'annotation. Ces règles formalisent des principes biologiques admis par la communauté scientifique. Elles se répartissent en trois classes de portée croissante : (i) les règles élémentaires ; (ii) les règles chromosomiques et (iii) les règles génomiques. Leur application en suivant cet ordre permet de détecter et donc de corriger d'abord les erreurs majeures. Un sous-ensemble de ces règles est mis en œuvre lors du processus d'annotation pour le projet Génolevures. Ceci a permis la détection d'erreurs portant par exemple sur la structure génique, la syntaxe ou bien l'absence de gènes essentiels. Certaines règles sont partie intégrante du système d'annotation semi-automatique MAGUS, utilisé pour la phase Génolevures 3. Cela permet de détecter au plus tôt les erreurs. Ce système fait suite au premier système qui était basé sur le logiciel CAAT-Box, et que nous avions mis en place pour l'annotation des levures de la phase Génolevures 2. Ce premier système d'annotation manuelle permettait de mieux gérer et donc d'accroître l'efficacité des annotateurs, mais aussi de garantir l'homogénéité de l'annotation réalisée pour les levures *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*.

Par ailleurs, l'étude expérimentale des interactions protéine-protéine par électrophorèse BN/SDS chez la levure *Y. lipolytica* a apporté la preuve biologique de la qualité de la straté-

gie d'annotation adoptée par le consortium Génolevures concernant la détection des CDS et des introns. La mise en évidence de complexes protéiques chez *S. cerevisiae* et *Y. lipolytica*, précédemment identifiés en partie chez *S. cerevisiae* ou prédits chez *Y. lipolytica* par homologie avec *S. cerevisiae*, permet d'enrichir les connaissances existantes sur ces levures. Les interactions protéine-protéine ainsi identifiées permettent notamment de confirmer les annotations fonctionnelles et relationnelles réalisées mais aussi d'étayer ces dernières. De plus, ces résultats fournissent un socle de connaissances prouvées, car d'origine expérimentale, pour mener de nouvelles analyses.

Le travail réalisé dans les domaines de la bio-informatique et de la biologie cellulaire ouvre ainsi de nouvelles perspectives. Les règles de cohérence sont un moyen désormais indispensable à l'annotation d'un génome afin d'assurer sa qualité. Le développement de certaines règles, telles celles qui concernent la cohérence de l'annotation pour les gènes codant des protéines appartenant à une même famille, rejoint d'autres problématiques de recherche portant sur le "text mining" et l'utilisation de vocabulaire contrôlé. Par ailleurs, ces règles ont été développées de façon générique, ce qui leur confère la propriété d'être applicables à tout organisme. Notre approche permet d'intégrer des règles spécialisées pour s'adapter à d'autres organismes et à d'autres analyses disponibles pour l'annotation. Des règles plus fines pourraient également être appliquées pour prendre en compte par exemple les sites de liaisons de facteurs de transcription dans le promoteur d'un gène, ou bien la présence du site terminateur de la transcription en aval de la séquence codante.

L'intégration de l'ensemble des règles définies dans ces travaux de thèse à tout projet d'annotation génomique, permettrait d'obtenir des annotations de meilleure qualité, ce gain étant supérieur dans le cadre d'une annotation automatique. Cette différence de qualité moindre entre génomes annotés manuellement et ceux annotés automatiquement, permettrait leur comparaison *in silico*. Ainsi, quelque soit le genre d'annotation, les scientifiques auraient une plus grande confiance dans les observations obtenues. Les différences génomiques mises en évidence traduiraient alors les spécificités évolutives et fonctionnelles de chacun des organismes. Néanmoins, les résultats *in silico*, aussi excellents qu'ils soient, ne peuvent se passer d'une validation expérimentale, car la Nature semble avoir toujours une exception à la règle...

Bibliographie

[Aalto et al., 1993] Aalto, M., Ronne, H., and Keränen, S. (1993). Yeast syntaxins Sso1p and Sso2p belong to a family of related membrane proteins that function in vesicular transport. *EMBO J*, 12(11) :4095–104.

[Adams et al., 2000] Adams, M. et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461) :2185–2195.

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 28–36.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, 4th edition.

[Alexandersson et al., 2003] Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM : Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, 13(3) :496–502.

[Allen et al., 2006] Allen, J., Majoros, W., Pertea, M., and Salzberg, S. (2006). JIGSAW, GeneZilla and GlimmerHMM : puzzling out the features of human genes in the ENCODE regions. *Genome Biology*, 7(Suppl. 1) :S9.1–13.

[Allen and Salzberg, 2005] Allen, J. and Salzberg, S. (2005). JIGSAW : integration of multiple sources of evidence for gene prediction. *Bioinformatics*, 21(18) :3596–3603.

[Altschul et al., 1990] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410.

[Angellotti et al., 2007] Angellotti, M., Bhuiyan, S., Chen, G., and Wan, X. (2007). CodonO : codon usage bias analysis within and across genomes. *Nucleic Acids Research*, 35(Web Server Issue) :W132–W136.

[Apweiler et al., 2004] Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., and Yeh, L. (2004). UniProt : the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue) :D115–D119.

[Apweiler et al., 2007] Apweiler, R., Bairoch, A., and Wu, C. T. U. C. (2007). The universal protein resource (UniProt). *Nucleic Acids Research*, 35(Database Issue) :D193–197.

[Artamonova et al., 2007] Artamonova, I., Frishman, G., and Frishman, D. (2007). Applying negative rule mining to improve genome annotation. *BMC Bioinformatics*, 8(261).

[Ashburner et al., 2000] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25 :25–29.

[Bailey and Elkan, 1994] Bailey, T. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In AAAI Press, Menlo Park, C., editor, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36.

[Bajic et al., 2006] Bajic, V., Brent, M., Brown, R., Frankish, A., Harrow, J., Ohler, U., Solovyev, V., and Tan, S. (2006). Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biology*, 7(Suppl. 1) :S3.

[Baldi et al., 2000] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics*, 16(5) :412–424.

[Barth and Gaillardin, 1996] Barth, G. and Gaillardin, C. (1996). *Non-conventional Yeast in Biotechnology : a Handbook*. Berlin : Springer,Wolf, K.

[Batt, 2001] Batt, G. (2001). Représentation des interactions protéine/protéine dans la cadre d'une méthode de modélisation de réseaux géniques. INRIA, Rapport de Recherche.

[Beckerich et al., 1998] Beckerich, J., Boisramé-Baudevin, A., and Gaillardin, C. (1998). *Yarrowia lipolytica* : a model organism for protein secretion studies. *International Microbioly*, 1 :123–130.

[Berger et al., 1996] Berger, J., Gamblin, S., Harrison, S., and Wang, J. (1996). Structure and mechanism of DNA topoisomerase II. *Nature*, 379(6562) :225–232.

[Berget et al., 1977] Berget, S., Moore, C., and Sharp, P. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences U.S.A.*, 74(8) :3171–3175.

[Berks and the *C. elegans* GMSC, 1995] Berks, M. and the *C. elegans* GMSC (1995). The *C. elegans* genome sequencing project. *Genome Research*, 5(2) :99–104. GMSC : Genome Mapping and Sequencing Consortium.

[Bernardi, 1997] Bernardi, G., editor (1997). *Junk* DNA *: The Role and the Evolution of Non-coding Sequences*, volume 205. Elsevier.Gene.

[Beyne et al., 2005] Beyne, E., Lasserre, J., Claverol, S., Sherman, D., and Bonneu, M. (2005). Identification and comparison of yeast complexomes. ESF-EMBO Symposium Comparative Genomics of Eukaroytic Microorganisms, San Feliu de Guixols, Spain.

[Biswas et al., 2002] Biswas, M., O'Rourke, J., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F., and Apweiler, R. (2002). Applications of InterPro in protein annotation and genome analysis. *Brief Bioinformatics*, 3(3) :285–295.

[Blandin et al., 2000] Blandin, G., Durrens, P., Tekaia, F., Aigle, M., Bolotin-Fukuhara, M., Bon, E., Casarégola, S., de Montigny, J., Gaillardin, C., Lépingle, A., Llorente, B., Malpertuy, A., Neuvéglise, C., Ozier-Kalogeropoulos, O., Perrin, A., Potier, S., Souciet, J., Talla, E., Toffano-Nioche, C., Wésolowski-Louvel, M., Marck, C., and Dujon, B. (2000). Genomic exploration of the hemiascomycetous yeasts : 4. the genome of *Saccharomyces cerevisiae* revisited. FEBS *Letters Special Issue*, 487(1) :31–36.

[Blattner et al., 1997] Blattner, F., Plunkett 3rd, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* k-12. *Science*, 277(5331) :1453–1474.

[Bon et al., 2003] Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M., Guldener, U., Mewes, H., Van Helden, J., Dujon, B., and Gaillardin, C. (2003). Molecular evolution of eukaryotic genomes : hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Research*, 31(4) :1121–1135.

[Borodovsky and McIninch, 1993] Borodovsky, M. and McIninch, J. (1993). GeneMark : parallel gene recognition for both DNA strands. *Computers & Chemistry*, 17(19) :123–133. http ://opal.biology.gatech.edu/GeneMark.

[Borodovsky et al., 1995] Borodovsky, M., McIninch, J., Koonin, E., Rudd, K., Médigue, C., and Danchin, A. (1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Research*, 23(17) :3554–3562.

[Brent, 2005] Brent, M. (2005). Genome annotation past, present, and future : How to define an orf at each locus. *Genome Res*, 15(12) :1777–1786.

[Brosius and Gould, 1992] Brosius, J. and Gould, S. (1992). On "genomenclature" : a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proceedings of the National Academy of Sciences U.S.A.*, 89(22) :10706–10710.

[Brown, 2002] Brown, T. (2002). *Genomes*. BIOS Scientific Publishers Ltd., 2nd. edition.

[Bryson et al., 2006] Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., van de Guchte, M., Penaud, S., Maguin, E., Hoebeke, M., Bessières, P., and Gibrat, J. (2006). AGMIAL : implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Research*, 34(12) :3533–3545.

[Bulmer, 1987] Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature*, 325(6106) :728–730.

[Burge and Karlin, 1997] Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1) :78–94.

[Camacho-Carvajal et al., 2004] Camacho-Carvajal, M., Wollscheid, B., Aebersold, R., Steimle, V., and Schamel, W. (2004). Two-dimensional blue native/SDS gel electrophoresis of multi-protein complexes from whole cellular lysates : a proteomics approach. *Molecular & Cellular Proteomics*, 3(2) :176–182.

[Casaregola et al., 2000] Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C., Artiguenave, F., Wincker, P., and Gaillardin, C. (2000). Genomic exploration of the hemiascomycetous yeasts : 17. *Yarrowia lipolytica*. FEBS *Letters Special Issue*, 487(1) :95–100.

[Castillo-Davis, 2005] Castillo-Davis, C. (2005). The evolution of noncoding DNA : how much junk, how much func ? *Trends in Genetics*, 21(10) :533–536.

[Chaudhri et al., 2003] Chaudhri, M., Scarabel, M., and Aitken, A. (2003). Mammalian and yeast 14-3-3 isoforms form distinct patterns of dimers *in vivo*. *Biochem Biophys Res Commun*, 300(3) :679–85.

[Chen et al., 2006] Chen, J., Hsu, W., Lee, M., and Ng, S. (2006). Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, 22(16) :1998–2004.

[Cherry et al., 1998] Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD : *Saccharomyces Genome* Database. *Nucleic Acids Research*, 26(1) :73–80.

[Chow et al., 1977] Chow, L., Gelinas, R., Broker, T., and Roberts, R. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1) :1–8.

[Coissac et al., 1997] Coissac, E., Maillier, E., and Netter, P. (1997). A comparative study of duplications in bacteria and eukaryotes : the importance of telomeres. *Mol Biol Evol*, 14(10) :1062–1074.

[consortium, 2005] consortium, H. (2005). Proteomics' new order. *Nature*, 437(7056) :169–170. Editorial.

[Consortium, 2001] Consortium, T. G. O. (2001). Creating the Gene Ontology resource : Design and implementation. *Genome Research*, 11 :1425–1433.

[Cooper, 2000] Cooper, G. (2000). *The Cell : a molecular approach*. Sinauer Associates, Inc., 2nd. edition. Boston University.

[Coriton et al., 2000] Coriton, O., Lepourcelet, M., Hampe, A., Galibert, F., and Mosser, J. (2000). Transcriptional analysis of the 69-kb sequence centromeric to HLA-J : a dense and complex structure of five genes. *Mammalian Genome*, 11(12) :1127–1131.

[Cotter et al., 2004] Cotter, D., Guda, P., Fahy, E., and Subramaniam, S. (2004). MitoProteome : mitochondrial protein sequence database and annotation system. *Nucleic Acids Research*, 32(Database issue) :D463–D467.

[Cumsky et al., 1987] Cumsky, M., Trueblood, C., Ko, C., and Poyton, R. (1987). Structural analysis of two genes encoding divergent forms of yeast cytochrome c oxidase subunit v. *Molecular Cell Biology*, 7(10) :3511–3519.

[David et al., 2006] David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C., Bofkin, L., Jones, T., Davis, R., and Steinmetz, L. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Acadademy of Sciences U.S.A.*, 103(14) :5320–5325.

[De Hertog et al., 2003] De Hertog, B., Talla, E., Tekaia, F., Beyne, E., Sherman, D., and Baret, P. (2003). Novel transporters from hemoascomycete yeasts. *Journal of Molecular Microbiology and Biotechnology*, 6(1) :19–28.

[De Hertogh et al., 2002] De Hertogh, B., Carvajal, E., Talla, E., Dujon, B., Baret, P., and Goffeau, A. (2002). Phylogenetic classification of transporters and other membrane proteins from *Saccharomyces cerevisiae*. *Functional Integrative Genomics*, 2(4-5) :154–170.

[Dietrich et al., 2004] Dietrich, F., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., Wing, R., Flavier, A., Gaffney, T., and Philippsen, P. (2004). The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, 304(5668) :304–307.

[Dimitri et al., 2005] Dimitri, P., Corradini, N., Rossi, F., and Verni, F. (2005). The paradox of functional heterochromatin. *Bioessays*, 27(1) :29–41.

[Djebali et al., 2006] Djebali, S., Delaplace, F., and Crollius, H. (2006). Exogean : a framework for annotating protein-coding genes in eukaryotic genomic DNA. *Genome Biology*, 7(Suppl. 1) :S7.

[Do and Choi, 2006] Do, J. and Choi, D. (2006). Computational approaches to gene prediction. *Journal of Microbiology*, 44(2) :137–144.

[Dujon et al., 1994] Dujon, B., Alexandraki, D., André, B., Ansorge, W., Baladron, V., Ballesta, J., Banrevi, A., Bolle, P., Bolotin-Fukuhara, M., and Bossier, P. (1994). Complete DNA sequence of yeast chromosome XI. *Nature*, 369(6479) :371–378.

[Dujon et al., 2004] Dujon, B., Sherman, D., et al. (2004). Genome evolution in yeasts. *Nature*, 430(6995) :35–44.

[Durbin and Thierry-Mieg, 1991] Durbin, R. and Thierry-Mieg, J. (1991). ACeDB. http ://www.acedb.org.

[Eilbeck et al., 2005] Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology : a tool for the unification of genome annotations. *Genome Biology*, 6(5) :R44.

[*C. elegans* Sequencing Consortium, 1998] *C. elegans* Sequencing Consortium, T. (1998). Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science*, 282(5396) :2012–2018.

[Engelhardt et al., 2005] Engelhardt, B., Jordan, M., Muratore, K., and Brenner, S. (2005). Protein molecular function prediction by bayesian phylogenomics. *PLoS Computational Biology*, 1(5) :e45.

[Etzold et al., 1996] Etzold, T., Ulyanov, A., and Argos, P. (1996). SRS : information retrieval system for molecular biology data banks. *Methods Enzymol*, 266 :114–128.

[Feldmann, 2005] Feldmann, H. (2005). *Yeast molecular biology : a short compedium on basis features and novels apsects*. University of Munich. http ://biochemie.web.med.uni-muenchen.de/Yeast_Biol/index.htm.

[Fichant and Burks, 1991] Fichant, G. and Burks, C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology*, 220(3) :659–671.

[Fickett, 1996] Fickett, J. (1996). Finding genes by computer : the state of the art. *Trends in Genetics*, 12(8) :316–320.

[Fields and Song, 1989] Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340 :245–246.

[Fitch, 2000] Fitch, W. (2000). Homology : a personal view on some of the problems. *Trends in Genetics*, 16(5) :227–231.

[Fleischmann et al., 1995] Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., and Merrick, J. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd. Science*, 269(5223) :496–512.

[Frangeul et al., 2004] Frangeul, L., Glaser, P., Rusniok, C., Buchrieser, C., Duchaud, E., Dehoux, P., and Kunst, F. (2004). CAAT-Box, contigs-assembly and annotation tool-box for genome sequencing projects. *Bioinformatics*, 20(5) :790–797.

[Fraser et al., 1995] Fraser, C., Gocayne, J., et al. (1995). The minimal gene complement of *Mycoplasma genitalium. Science*, 270(5235) :397–403.

[Frishman et al., 2001] Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., and Mewes, H. (2001). Functional and structural genomics using PEDANT. *Bioinformatics*, 17(1) :44–57.

[Gaasterland et al., 2000] Gaasterland, T., Sczyrba, A., Thomas, E., Aytekin-Kurban, G., Gordon, P., and Sensen, C. (2000). MAGPIE/EGRET annotation of the 2.9-Mb *Drosophila melanogaster* Adh region. *Genome Research*, 10(4) :502–510.

[Gaasterland and Sensen, 1996a] Gaasterland, T. and Sensen, C. (1996a). Fully automated genome analysis that reflects user needs and preferences. a detailed introduction to the MAGPIE system architecture. *Biochimie*, 78(5) :302–310.

[Gaasterland and Sensen, 1996b] Gaasterland, T. and Sensen, C. (1996b). MAGPIE : automated genome interpretation. *Trends in Genetics*, 12(2) :76–78.

[Gabaldon et al., 2007] Gabaldon, T., Pereto, J., Montero, F., Gil, R.and Latorre, A., and Moya, A. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical transactions of the Royal Society of London, Series B, Biological sciences*, pages 1751–1762.

[Gallwitz and Sures, 1980] Gallwitz, D. and Sures, I. (1980). Structure of a split yeast gene : complete nucleotide sequence of the actin gene in *Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences U.S.A.*, 77(5) :2546–50.

[Galperin and Koonin, 1998] Galperin, M. and Koonin, E. (1998). Sources of systematic error in functional annotation of genomes : domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology*, 1(1) :55–67.

[Gattiker et al., 2002] Gattiker, A., Gasteiger, E., and Bairoch, A. (2002). ScanProsite : a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*, 1(2) :107–108.

[Gavin et al., 2002] Gavin, A. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415 :141–147.

[Giaever et al., 2002] Giaever, G. et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896) :387–391.

[Gish, 1995] Gish, W. (1995). http ://blast.wustl.edu/.

[Glover and Hogness, 1977] Glover, D. and Hogness, D. (1977). A novel arrangement of the 18S and 28S sequences in a repeating unit of *Drosophila melanogaster* rDNA. *Cell*, 10(2) :167–176.

[Goff et al., 2002] Goff, S. et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science*, 296(5565) :92–100.

[Goffeau et al., 1996] Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., et al. (1996). Life with 6 000 genes. *Science*, 274(5287) :546,563–567.

[Grantham et al., 1980] Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1) :49–62.

[Green and Hillier, ults] Green, P. and Hillier, L. (unpublished results). GeneFinder.

[Griffiths et al., 2002] Griffiths, A. et al. (2002). *Modern Genetic analysis : integrating genes and genomes*, page 299. Freeman, W.H. and Compagny, second edition.

[Guigó and Fickett, 1995] Guigó, R. and Fickett, J. (1995). Distinctive sequence features in protein coding genic non-coding, and intergenic Human DNA. *Journal of Molecular Biology*, 253(1) :51–60.

[Guigo et al., 2006] Guigo, R., Flicek, P., Abril, J., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T., Harrow, J., Hubbard, T., Lewis, S., and Reese, M. (2006). EGASP : the human ENCODE genome annotation assessment project. *Genome Biology*, 7(Suppl 1) :S2.1–31.

[Harrow et al., 2006] Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., and Guigo, R. (2006). GENCODE : producing a reference annotation for ENCODE. *Genome Biology*, 7(Suppl. 1).

[Hermjakob et al., 2004a] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., Von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004a). The HUPO PSI's Molecular Interaction format-a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2) :177–83.

[Hermjakob et al., 2004b] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004b). IntAct - an open source molecular interaction database. *Nucleic Acids Research*, 32(Database issue) :D452–D455.

[Higgins et al., 1994] Higgins, D., Thompson, J., Gibson, T., Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W : improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22 :4673–4680.

[Ho et al., 2002] Ho, Y. et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868) :180–183.

[Hoeffken et al., 1988] Hoeffken, H., Siegward, H., Bartlett, P., and Huber, R. (1988). Crystal structure determination, refinement and molecular model of creatine amidinohydrolase from *Pseudomonas putida*. *Journal of Molecular Biology*, 204 :417–433.

[Hurt, 1988] Hurt, E. (1988). A novel nucleoskeletal-like protein located at the nuclear periphery is required for the life cycle of *Saccharomyces cerevisiae*. *EMBO Journal*, 7(13) :4323–4334.

[Initiative, 2000] Initiative, T. A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814) :796–815.

[Iragne, 2006] Iragne, F. (2006). Pathways conservation. http ://cbi.labri.fr/Genolevures/other_studies.php.

[Iragne et al., 2007] Iragne, F., Nikolski, M., and Sherman, D. (2007). Extrapolation of metabolic pathways as an aid to modelling completely sequenced nonsaccharomyces yeasts. *FEMS Yeast Research*.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Acadademy of Sciences U.S.A.*, 98(8) :4569–4574.

[Jansen et al., 2003] Jansen, R., Bussemaker, H., and Gerstein, M. (2003). Revisiting the codon adaptation index from a whole-genome perspective : analyzing the relationship between gene expression and codon occurrence in yeast using a variety of model. *Nucleic Acids Research*, 31(8) :2242–2251.

[Johnston et al., 1994] Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Du, Z., Favello, A., Fulton, L., and Gattung, S. (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science*, 265(5181) :2077–2082.

[Jones et al., 2007] Jones, C., Brown, A., and Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8(170) :364–373.

[Kanehisa et al., 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics : new developments in KEGG. *Nucleic Acids Research*, 34(Database Issue) :D354–D357.

[Karlin et al., 2003] Karlin, S., Mrázek, J., and Gentles, A. (2003). Genome comparisons and analysis. *Curr Opin Struct Biol*, 13(3) :344–352.

[Kellis et al., 2004] Kellis, M., Birren, B., and Lander, E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983) :617–624.

[Kerrien et al., 2007] Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A., Vinod, N., Bader, G., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J., Moore, S., Ceol, A., Chatr-Aryamontri, A., Oesterheld, M., Stumpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007). Broadening the

horizon - level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5(1) :44.

[Kessler et al., 2003] Kessler, M., Zeng, Q., Hogan, S., Cook, R., Morales, A., and Cottarel, G. (2003). Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Research*, 13(2) :264–271.

[Kretschmann et al., 2001] Kretschmann, E., Fleischmann, W., and Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17(10) :920–926.

[Krogan et al., 2004] Krogan, N., Peng, W., Cagney, G., Robinson, M., Haw, R., Zhong, G., Guo, X., Zhang, X., Canadien, V., Richards, D., Beattie, B., Lalev, A., Zhang, W., Davierwala, A., Mnaimneh, S., Starostine, A., Tikuisis, A., Grigull, J., Datta, N., Bray, J., Hughes, T., Emili, A., and Greenblatt, J. (2004). High-definition macromolecular composition of yeast RNA-processing complexes. *Mol Cell*, 13(2) :225–39.

[Lafontaine et al., 2004] Lafontaine, I., Fischer, G., Talla, E., and Dujon, B. (2004). Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene*, 335.

[Lasserre et al., 2005] Lasserre, J., Beyne, E., Claverol, S., Sherman, D., and Bonneu, M. (2005). Nouveau procédé pour l'analyse de l'ensemble des complexes protéiques d'une cellule. SFEAP 2005 I^{er}Symposium de Chimie et Biologie Analytiques, Montpellier, France.

[Lasserre et al., 2006] Lasserre, J., Beyne, E., Pyndiah, S., Lapaillerie, D., Claverol, S., and Bonneu, M. (2006). A complexomic study of *Escherichia coli* using two-dimensional blue native/SDS polyacrylamide gel electrophoresis. *Electrophoresis*, 27(16) :3306–3321.

[Legrain et al., 2001] Legrain, P., Wojcik, J., and Gauthier, J. (2001). Protein–protein interaction maps : a lead towards cellular functions. *Trends in Genetics*, 17(6) :346–352.

[Lewis et al., 2002] Lewis, S., Searle, S., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M., Kaminker, J., Matthews, B., Prochnik, S., Smithy, C., Tupy, J., Rubin, G., Misra, S., Mungall, C., and Clamp, M. (2002). Apollo : a sequence annotation editor. *Genome Biol*, 3(12) :RESEARCH0082.

[Lowe and Eddy, 1997] Lowe, T. and Eddy, S. (1997). tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5) :955–964.

[Majoros et al., 2005] Majoros, W., Pertea, M., Delcher, A., and Salzberg, S. (2005). Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinformatics*, 6(1) :16.

[Majoros et al., 2004] Majoros, W., Pertea, M., and Salzberg, S. (2004). TigrScan and GlimmerHMM : two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, 20(16) :2878–2879.

[Makalowski et al., 1996] Makalowski, W., Zhang, J., and Boguski, M. (1996). Comparative analysis of 1 196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Research*, 6(9) :846–857.

[Marcotte et al., 1999] Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428) :751–753.

[Marin et al., 2002] Marin, A., Pothier, J., Zimmermann, K., and Gibrat, J. (2002). FROST : a filter-based fold recognition method. *Proteins*, 49(4) :493–509.

[Mathé et al., 2002] Mathé, C., Sagot, M., Schiex, T., and Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19) :4103–4117.

[McClintock, 1953] McClintock, B. (1953). Induction of instability at selected loci in maize. *Genetics*, 38(6) :579–599.

[McLachlan et al., 1984] McLachlan, A., Staden, R., and Boswell, D. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Research*, 12(24) :9567–9575.

[Médigue et al., 1999] Médigue, C., Rechenmann, F., Danchin, A., and Viari, A. (1999). Imagene : an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, 15(1) :2–15.

[Müller et al., 1998] Müller, S., Sandal, T., Kamp-Hansen, P., and Dalboge, H. (1998). Comparison of expression systems in the yeasts *Saccharomyces cerevisiae*, *Hansenula polymorpha*, *Klyveromyces lactis*, *Schizosaccharomyces pombe* and *Yarrowia lipolytica*. Cloning of two novel promoters from *Yarrowia lipolytica*. *Yeast*, 14(14) :1267–1283.

[Nadir et al., 1996] Nadir, E., Margalit, H., Gallily, T., and Ben-Sasson, S. (1996). Microsatellite spreading in the human genome : evolutionary mechanisms and structural implications. *Proceedings of the National Academy of Sciences U.S.A.*, 93(13) :6470–6475.

[Nakase et al., 1998] Nakase, T., Suzuki, M., Phaff, H., and Kurtzman, C. (1998). *The Yeasts, a Taxonomic Study*. Elsevier, Amsterdam, 4th edition.

[Neuvéglise, 2005] Neuvéglise, C. (2005). Génosplicing. http ://cbi.labri.fr/Genolevures/genosplicing/index.php.

[Nicaud et al., 2002] Nicaud, J., Madzak, P., van den Broeck, P., Gysler, C., Duboc, P., Niederberger, P., and Gaillardin, C. (2002). Protein expression and secretion in the yeast *Yarrowia lipolytica*. *FEMS Yeast Research*, 2 :371–379.

[Nikolski and Sherman, 2007] Nikolski, M. and Sherman, D. (2007). Family relationships : should consensus reign ?-consensus clustering for protein families. *Bioinformatics*, 23(2) :e71–e76.

[Nooren and Thornton, 2003] Nooren, I. and Thornton, J. (2003). Diversity of protein-protein interactions. *EMBO Journal*, 22(14) :3486–3492.

[Notredame et al., 2000] Notredame, C., Higgins, D., and Heringa, J. (2000). T-coffee : a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1) :205–217.

[Oliver and Marín, 1996] Oliver, J. and Marín, A. (1996). A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*, 43(3) :216–23.

[Oliver et al., 1992] Oliver, S., van der Aart, Q., Agostoni-Carbone, M., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J., and Benit, P. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, 357 :38–46.

[Orchard et al., 2006] Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi1, L., Oesterheld, M., St.mpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J., Moore, S., Wojcik1, J., Bader, G., Vidal, M., Cusick, M., Gerstein, M., Gavin, A., Superti-Furga11, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Consortium, T. G., Gilson, M., Hogue, C., Mewes, H., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2006). The minimum information required for reporting a molecular interaction experiment (mimix). *Proteomics*, 6(16).

[Payseur and Nachman, 2002] Payseur, B. and Nachman, M. (2002). Gene density and human nucleotide polymorphism. *Molecular Biology Evolution*, 19(3) :336–340.

[Pearson and Lipman, 1988] Pearson, W. and Lipman, D. (1988). Improved tools for biological sequence comparison. *Procedure of the National Acadademy of Sciences U.S.A.*, 85(8) :2444–2448.

[Pellegrini et al., 1977] Pellegrini, M., Manning, J., and Davidson, N. (1977). Sequence arrangement of the rDNA of *Drosophila melanogaster*. *Cell*, 10(2) :213–224.

[Pellegrini et al., 1999] Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. (1999). Assigning protein functions by comparative genome analysis : protein phylogenetic profiles. *Proceedings of the National Academy of Sciences U.S.A.*, 96(8) :4285–4288.

[Percudani et al., 1997] Percudani, R., Pavesi, A., and Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 268(2) :322–330.

[Petsko, 2003] Petsko, G. (2003). Funky, not junky. *Genome Biology*, 4(2) :104.

[Pignede et al., 2000] Pignede, G., Wang, H., Fudalej, F., Gaillardin, C., Seman, M., and Nicaud, J. (2000). Characterization of an extracellular lipase encoded by LIP2 in *Yarrowia lipolytica*. *Journal of Bacteriology*, 182(10) :2802–2810.

[Puig et al., 2001] Puig, O., Caspary, F., Rigaut, G., et al. (2001). The tandem affinity purification (TAP) method : A general procedure of protein complex purification. *Methods*, 24 :218–29.

[Quilan, 1993] Quilan, J. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

[Reese et al., 2000] Reese, M., Hartzell, G., Harris, N., Ohler, U., Abril, J., and Lewis, S. (2000). Genome annotation assessment in *Drosophila melanogaster*. *Genome Research*, 10(4) :483–501.

[Rigaut et al., 1999] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17 :1030–1032.

[Riley et al., 2007] Riley, M., Schmidt, T., Artamonova, I., Wagner, C., Volz, A., Heumann, K., Mewes, H., and Frishman, D. (2007). PEDANT genome database : 10 years online. *Nucleic Acids Research*, 35(Database issue) :D354–D357.

[Ross et al., 2004] Ross, P., Huang, Y., Marchese, J., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics*, 3(12) :1154–1169.

[Rubin et al., 2000] Rubin, G., Yandell, M., et al. (2000). Comparative genomics of the eukaryotes. *Science*, 287(5461) :2204–2215.

[Rutherford et al., 2000] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., and Barrell, B. (2000). Artemis : sequence visualization and annotation. *Bioinformatics*, 16(10) :944–945.

[Saeys et al., 2007] Saeys, Y., Rouzé, P., and Van de Peer, Y. (2007). In search of the small ones : improved prediction of short exons in vertebrates, plants, fungi, and protists. *Bioinformatics*, 23(4) :414–420.

[Salzberg et al., 1998] Salzberg, S., Delcher, A., Kasif, S., and White, . (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2) :544–548.

[Sambrook and Russel, 2001] Sambrook, J. and Russel, D. (2001). *Molecular Cloning : a molecular manual*, volume 1. Cold Spring Harbor Laboratory, 3 edition.

[Schägger, 1995] Schägger, H. (1995). Quantification of oxidative phosphorylation enzymes after blue native electrophoresis and two-dimensional resolution : normal complex I protein amounts in Parkinson's disease conflict with reduced catalytic activities. *Electrophoresis*, 16 :763–770.

[Schägger, 2006] Schägger, H. (2006). Tricine-SDS-PAGE. *Nature Protocoles*, 1(1) :16–22.

[Schägger et al., 1996] Schägger, H., Bentlage, H., Ruitenbeek, W., Pfeiffer, K., Rotter, S., Rother, C., Böttcher-Purkl, A., and Lodemann, E. (1996). Electrophoretic separation of multiprotein complexes from blood platelets and cell lines : technique for the analysis of diseases with defects in oxidative phosphorylation. *Electrophoresis*, 17(4) :709–714.

[Schägger and von Jagow, 1987] Schägger, H. and von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analytical Biochemistry*, 166(2) :368–379.

[Schägger and von Jagow, 1991] Schägger, H. and von Jagow, G. (1991). Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Analytical Biochemistry*, 199 :223–231.

[Serebriiskii and Golemis, 2001] Serebriiskii, I. and Golemis, E. (2001). Two-hybrid system and false positives. Approaches to detection and elimination. *Methods in Molecular Biology*, 177 :123–134.

[Sherman et al., 2004] Sherman, D., Durrens, P., Beyne, E., Nikolski, M., and Souciet, J. G. C. (2004). Génolevures : comparative genomics and molecular evolution of hemiascomycetous yeasts. *ucleic Acids Research*, 32(Database Issue) :315–318.

[Sherman et al., 2006] Sherman, D., Durrens, P., Iragne, F., Beyne, E., Nikolski, M., and Souciet, J. (2006). Génolevures complete genomes provide data and tools for comparative

genomics of hemiascomycetous yeast. *Nucleic Acids Research*, 34(Database issue) :D432–435.

[Sigrist et al., 2002] Sigrist, C., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). PROSITE : a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3 :265–274.

[Smalley et al., 2003] Smalley, D. et al. (2003). In search of the minimal *Escherichia coli* genome. *Trends in Microbiology*, 11(1) :6–8.

[Smith and Waterman, 1981] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 :195–197.

[Sonnhammer et al., 1997] Sonnhammer, E., Eddy, S., and Durbin, R. (1997). Pfam : a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3) :405–420.

[Sorimachi et al., 1991] Sorimachi, H., Kawasaki, H., Tsukahara, T., Ishiura, S., Emori, Y., Sugita, H., and Suzuki, K. (1991). Sequence comparison among subunits of multicatalytic proteinase. *Biomed Biochim Acta*, 50 :459–564.

[Souciet et al., 2000] Souciet, J. et al. (2000). Génolevures : Genomic exploration of the hemiascomycetous yeasts. FEBS *Letters Special Issue*, 487(1). http ://www.elsevier.nl/febs/125/18/show/toc.htt.

[Sprinzak et al., 2003] Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data ? *Journal of Molecular Biology*, 327(5) :919–923.

[Stanke, 2003] Stanke, M. (2003). *Gene Prediction with a Hidden* M*arkov Model*. PhD thesis, Georg-August-Universität Göttingen.

[Stark et al., 2006] Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID : a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue) :D535–53–9. http ://www.thebiogrid.org/.

[Strange, 2006] Strange, K. (2006). An overview of *C. elegans* biology. *Methods Mol Biol*, 351 :1–12.

[Strick et al., 1992] Strick, C., James, L., O'Donnell, M., Gollaher, M., and Franke, A. (1992). The isolation and characterization of the pyruvate kinase-encoding gene from the yeast *Yarrowia lipolytica*. *Gene*, 118 :65–72.

[Strick et al., 1994] Strick, C., James, L., O'Donnell, M., Gollaher, M., and Franke, A. (1994). The isolation and characterization of the pyruvate kinase-encoding gene from the yeast *Yarrowia lipolytica*. *Gene*, 140(1) :141–3.

[Sun et al., 2005] Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., Shi, T., and Li, Y. (2005). Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics*, 21(16) :3409–3415.

[Tatusov et al., 2003] Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Smirnov, S.,

Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J., and Natale, D. (2003). The COG database : an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1) :41.

[Tatusov et al., 2000] Tatusov, R., Galperin, M., Natale, D., and Koonin, E. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1) :33–36.

[Tatusov et al., 1997] Tatusov, R., Koonin, E., and Lipman, D. (1997). A genomic perspective on protein families. *Science*, 278(5338) :631–637.

[ENCODE Project consortium, 2004] ENCODE Project consortium (2004). The ENCODE (encyclopedia of dna elements) project. *Science*, 306 :636–640.

[Thierry-Mieg and Thierry-Mieg, 2006] Thierry-Mieg, D. and Thierry-Mieg, J. (2006). AceView : a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biology*, 7(Suppl. 1) :S12.1–14.

[Uetz et al., 2000] Uetz, P. et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770) :623–627.

[Valencia and Pazos, 2002] Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3) :368–373.

[Vallenet et al., 2006] Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Médigue, C. (2006). MaGe : a microbial genome annotation system supported by synteny results. *Nucleic Acids Research*, 34(1) :53–65.

[Velculescu et al., 1997] Velculescu, V., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M., Bassett Jr, D., Hieter, P., Vogelstein, B., and Kinzler, K. (1997). Characterization of the yeast transcriptome. *Cell*, 88(2) :243–251.

[Venter et al., 2001] Venter, J. et al. (2001). The sequence of the human genome. *Science*, 291(5507) :1304–1351.

[Vert, 2002] Vert, J. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18(Suppl. 1) :S276–284.

[Villa et al., 1996] Villa, A., Strina, D., Frattini, A., Faranda, S., Macchi, P., Finelli, P., Bozzi, F., Susani, L., Archidiacono, N., Rocchi, M., and Vezzoni, P. (1996). The ZNF75 zinc finger gene subfamily : isolation and mapping of the four members in humans and great apes. *Genomics*, 35(2) :312–320.

[von Mering et al., 2002] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887) :399–403.

[Wang et al., 2004] Wang, Z., Chen, Y., and Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics*, 2 :216–221.

[Warbrick, 1997] Warbrick, E. (1997). Tow's company, three's a crowd : the yeast two hybrid system for mapping molecular interactions. *Structure*, 5 :13–17.

[Waskiewicz-Staniorowska et al., 1998] Waskiewicz-Staniorowska, B., Skala, J., Jasinski, M., Grenson, M., Goffeau, A., and Ulaszewski, S. (1998). Functional analysis of three adjacent open reading frames from the right arm of yeast chromosome XVI. *Yeast*, 14 :1027–1039.

[Waterston et al., 2002] Waterston, R. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915) :520–562.

[Watson and Crick, 1953a] Watson, J. and Crick, F. (1953a). Genetical implications of the structure of desoxyribonucleic acid. *Nature*, 171(4361) :964–967.

[Watson and Crick, 1953b] Watson, J. and Crick, F. (1953b). Molecular structure of nucleic acids : a structure for desoxyribose nucleic acid. *Nature*, 171(4356) :737–738.

[Wellauer and Dawid, 1977] Wellauer, P. and Dawid, I. (1977). The structural organization of ribosomal DNA in *Drosophila melanogaster*. *Cell*, 10(2) :193–212.

[Wieser et al., 2004] Wieser, D., Kretschmann, E., and Apweiler, R. (2004). Filtering erroneous protein annotation. *Bioinformatics*, 20(Suppl 1) :i342–i347.

[Wittig et al., 2006] Wittig, I., Braun, H., and Schägger, H. (2006). Blue native PAGE. *Nature Protocoles*, 1(1) :418–428.

[Woese et al., 1990] Woese, C., Kandler, O., and Wheelis, M. (1990). Towards a natural system of organisms : Proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences U.S.A.*, 87 :4576–4579.

[Woese et al., 1978] Woese, C., Magrum, L., and Fox, G. (1978). Archaebacteria. *Journal of Molecular Evolution*, 11(3) :245–252.

[Wyman et al., 2004] Wyman, S., Jansen, R., and Boore, J. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, 20(17) :3252–3255.

[Yeo et al., 2004] Yeo, G., Hoon, S., Venkatesh, B., and Burge, C. (2004). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proceedings of the National Academy of Science U.S.A.*, 101(44) :15700–15705.

[Yu et al., 2002] Yu, J. et al. (2002). A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science*, 296(5565) :79–92.

[Zhang et al., 2004] Zhang, R., Ou, H., and Zhang, C. (2004). DEG : a database of essential genes. *Nucleic Acids Research*, 32(1) :D271–D272.

[Zheng and Gerstein, 2006] Zheng, D. and Gerstein, M. (2006). A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biology*, 7(Suppl. 1) :S13.

[Zuckerkandl, 1997] Zuckerkandl, E. (1997). Junk DNA and sectorial gene repression. *Gene*, 205(1-2) :323–343.

Liste des publications

[Lasserre et al., 2006] Lasserre, J.P., **Beyne, E.**, Pyndiah, S., Lapaillerie, D., Claverol, S. and Bonneu, M.. A complexomic study of *Escherichia coli* using two dimensional Blue Native / SDS Polyacrylamide Gel Electrophoresis. *Electrophoresis*, vol. 27, n^{o}16, August 2006, pp3306-21 .

[Sherman et al., 2006] Sherman, D., Durrens, P., Iragne, F., **Beyne, E.**, Nikolski, M. and Souciet, J.L.. Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeast. *Nucleic Acids Research*, vol. 34 (Database issue), January 2006, pp D432-5.

[Dujon et al., 2004] Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., **Beyne, E.**, Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P. and Souciet, J.L.. Genome Evolution in Yeasts. *Nature*, vol. 430, n^{o}6995, July 2004, pp 35-44.

[Sherman et al., 2004] Sherman, D., Durrens, P., **Beyne, E.**, Nikolski, M. and Souciet, J.L. ; Génolevures Consortium. Génolevures : comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Research*, vol. 32 (Database issue), January 2004, pp.D315-8.

[De Hertog et al., 2003] De Hertog, B., Talla, E., Tekaia, F., **Beyne, E.**, Sherman, D., Baret, P.V., Dujon, B. and Goffeau, A.. Novel transporters from hemoascomycete yeasts. *J. Mol. Microbiol. Biotechnol.*, vol. 6, n^{o}1, 2003, pp19-28.