Université de Montréal

Principes de l'évolution du réseau de l'homéostasie des protéines

par
Yasmine Draceni

Département de biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en bio-informatique

Décembre, 2018

Résumé

L'homéostasie cellulaire est la capacité d'une cellule à maintenir son équilibre et sa fonctionnalité. Une des causes de l'instabilité de cet équilibre est le stress. En effet, le stress provoque une accumulation des protéines mal repliées, qui peuvent former des agrégats provoquant des maladies neurodégénératives. Les protéines « chaperons » sont le principal mécanisme du repliement des protéines et du contrôle de leur qualité. Celles-ci forment le cœur d'un réseau qu'on appelle le réseau d'homéostasie des protéines. Celui-ci a pour but de contrôler, d'assurer et de protéger le protéome[1], par le biais de la réparation du repliement et l'élimination des agrégats. Le réseau joue un rôle essentiel pour garder l'homéostasie protéique cellulaire dite protéostasie. Actuellement, nous manquons de connaissances fondamentales sur la façon dont ce réseau fonctionne en équilibre, mais aussi comment il échoue lors d'un déséquilibre. Par exemple, le rat-taupe nu, *Heterocephalus glaber*, a un cycle de vie lent qui peut dépasser 30 ans. Il a un mécanisme résistant au stress et un bon système du réseau, ce qui lui permet d'atteindre facilement un équilibre au niveau de son fonctionnement cellulaire. À l'inverse, le poisson, *Nothobranchius furzeri*, dont l'espérance de vie est très courte présente un processus de vieillissement accéléré et une perturbation de l'homéostasie. Il est à propos de se demander, comment cet équilibre fonctionne-t-il chez ces organismes et chez d'autres ?

Ce projet de recherche utilise des approches bio-informatiques et de génomique comparative, afin de mettre en évidence les principes fondamentaux de la protéostasie. L'évolution du réseau des chaperons sera analysé dans le contexte de l'adaptation du protéome. En reliant à l'échelle évolutive, nous analyserons la diversification du réseau des chaperons à travers la phylogénie et l'unirons à l'évolution du protéome. Ce projet de maîtrise apporte des notions fondamentales sur l'évolution du réseau de l'homéostasie des protéines. Précisément, nous présentons une analyse comparative des chaperons de 216 espèces eucaryotiques qui indiquent

[1] Ensemble des protéines exprimées dans une cellule.

que l'équilibre de la protéostasie peut être une élément clé pour expliquer la robustesse[2] de l'organisme.

Mots-clés : Évolution, chaperons, homéostasie, bio-informatique, phylogénétique.

[2] Capacité à maintenir les performances face aux perturbations.

Abstract

Cell homeostasis is the ability of a cell to maintain its balance and functionality. One of the causes of the instability of this balance is stress. Indeed, stress causes an accumulation of misfolded proteins, which can form aggregates causing neurodegenerative diseases. Molecular chaperones are the main cellular mechanism that promote protein folding and quality control. These form the heart of a network called protein homeostasis network. The purpose of this is to control, insure and protect the proteome through folding and elimination of aggregates. The network plays a vital role in keeping cellular protein in homeostasis known as proteostasis. Failure of protein homeostasis is linked to aging and aging-associated neurodegenerative diseases such as Alzheimer's and Parkinson's. Currently, our understanding how this network keeps the proteome in balance in health, and how it fails and causes diseases, remains incomplete. Different species offer striking examples. For instance, the naked-mole rate, *Heterocephalus glaber,* remarkable for its life expectancy of over 30 years, it has an effective stress-resistant mechanism and a good homeostasis, which allows it to easily achieve a balance and homeostasis. On the other hand, the killifish, *Nothobranchius furzeri*, whose life expectancy is very short, has an accelerated aging process and with pronounced loss of homeostasis. Here we seek to ask, how does this balance work for these and other organisms?

This research project will use bioinformatics and comparative genomics approaches to highlight the fundamental principles of proteostasis. The evolution of the chaperone network will be analyzed in the context of proteome adaptation. We analyze the diversification of the chaperone network across the eukaryotic phylogeny and compare it with aspects of the evolution of the corresponding proteomes. This master's project provide fundamental insights into the biology and the evolution of the protein homeostasis network. Moreover, we use comparative genomics analysis of chaperons from 216 eukaryotes species which indicates that the balance in proteostasis could be a key variable in explaining organismal robustness.

Keywords: Chaperons, evolution, phylogenetics, homeostasis, bioinformatics.

Table des matières

Liste des tableaux

Chapitre 2

Chapitre 3

Annexe

Liste des figures

Chapitre 2

Chapitre 3

Liste des sigles

AAA	Associées à diverses Activités cellulaires
ADN	Acide Désoxyribonucléique
a.u	Atomic Units
ARN	L'acide ribonucléique
ARNr	ARN ribosomique
CCT	Chaperonines contenant TCP-1
Clp	Protéase caséinolytique P
CFI	Canada Foundation for Innovation
D	Acide aspartique
E	Acide glutamique
ETS	Espaceur transcrit externe
FRQ-NT	Fonds de recherche du Québec – Nature et technologies
GTR	Généralisé réversible dans le temps
HAT	Histone acétyltransférase
HDAC	Histones désacétylases
HSF1	Facteur du choc thermique 1
HSP	Protéine de choc thermique
HSR	Réponse au choc thermique cytosolique
HMMs	Hidden Markov models
ITS	espaceur non transcrit
JC	Jukes-Cantor
M	Methionine
MESI	Ministère de l'Économie, des Sciences et de l'Innovation du Québec
MC	Modèles de Covariance
MCMC	Méthode d'échantillonnage de Monte-Carlo par chaines de Markov
MCMCMC	Metropolis-Coupled Markov chain Monte Carlo
mRNA	Messenger RNA
MH	Metropolis-Hastings
N	asparagine

NBD	Domaines de liaison nucléotides
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NTS	espaceur non transcrit
NJ	Neighbour Joining
NP-complets	Non Déterministe polynomial
PKA	Protéines kinases A
PKG	Protéines kinases G
Q	glutamine
RE	Réticulum endoplasmique
RHP	Réseau de l'homéostasie des protéines
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
TCP	Tailles complexe polypeptide
TRiC	Tailles complexe polypeptide du complexe d'anneau
TPR	Répétition tétratricopeptide
UPR	Réponse aux protéines non pliées
UniProt	Universal Protein resource
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
V	Valine

Liste des abréviations

AA	Acides aminés
Asp	Aspartate
C.gigas	Crassostrea gigas
Glu	Glutamine
H.glaber	Heterocephalus glaber
N.furzeri	Nothobranchius furzeri
S. purpuratus	Strongylocentrotus purpuratus
P	Probabilité
Val	Valine

Remerciements

Je tiens tout d'abord adresser mes remerciements et ma reconnaissance à mon directeur de recherche, Dr. **Sebastian Pechmann**, pour m'avoir accueilli dans son laboratoire et pour m'avoir fais confiance d'attribuer ce projet fascinant. À mon humble avis, Sebastian possède des qualités d'un grand scientifique, par son excellant savoir dans le domaine de la bio-informatique, son ouverture d'esprit et son audace en recherche. Sans oublier, son côté humain et sa compréhension envers ses étudiants lui crée une ambiance chaleureuse et studieuse parmi les membres de son équipe. Je suis très honorée d'avoir été supervisé par un chercher aussi talentueux, ambitieux et passionné. Grâce à lui, j'ai découvert ce domaine dans lequel j'ai pu acquérir de nouvelles habiletés et connaissances.

Je remercie mon parrain, Dr. **Stephen Michnick**, pour les rencontres et les suggestions de projet.

Je remercie Dre. **Julie Hussin** et Dr. **Serguei Chteinberg** pour l'intérêt porté à ce travail en acceptant d'être les membres du jury de ce mémoire.

Mes grands remerciements aux membres du laboratoire, particulièrement à Pedro Bordignon pour son aide précieuse et ses discussions/débats constructifs scientifique ou non, à Savandra Besse pour ses astuces en programmation, discussions et conseils sur le projet ainsi que pour les corrections qu'elle m'a suggéré lors de la rédaction de ce mémoire. Je remercie Musa Oezboyaci, Mélissa Nelson, Nazli Kocatuğ, Amruta Sahoo, Chris kooistra, Vithuya Shanmugalingam, Lila Salhi, Matt Sarrasin, Amira Yazidi, Louis Gendron et Shamim Hasan pour les bons moments passés, ce fut une agréable expérience grâce à vous tou(te)s. Également, mes remerciements aux membres de l'AÉBINUM, particulièrement à Saraï Mola et Tariq Daouda.

Je suis notamment reconnaissante aux deux doctorants Emmanuel Noutahi et Simon Laurin-Lemay pour leur aide et leur précieux conseils portés sur le sujet de la phylogénie. Aussi, un merci à Marie Pageau qui m'a fait confiance pour l'enseignement aux laboratoires de bio-informatique et m'a réalisé à quel point j'aimais partager le savoir, à Dr. Sebastian Lemieux pour ses conseils en enseignement.

Un spécial remerciements à l'Université de Montréal et NSERC Discovery Grant pour leur soutien financier. Également, Élaine Meunier, pour ses conseils, ses anecdotes québécoises et les fous rires autour de son calendrier, à Audrey Noël qui m'a aussi fait confiance pour présenter et représenter le programme de bio-informatique de l'UdeM pour les portes ouvertes.

Enfin, un immense merci à ma famille qui a toujours été présente quand j'en avais besoin. Ma sœur, Mounia, qui a su me réconforter et m'encourager. Ma mère pour son amour, son attitude positive et ses encouragements malgré la distance. Mon père de croire en moi dans tous ce que je fais, de partager son amour pour la science, de m'inculquer les valeurs nobles digne de lui et de m'avoir appris à entreprendre l'inconnu.

À mon papa, pour ses innovations.

À ma maman, pour ses inspirations.

À Mounia.

Chapitre 1

Introduction

1.1 Mise en contexte

L'homéostasie cellulaire est la capacité d'une cellule à maintenir son équilibre. Une des causes de l'instabilité de cet équilibre est le stress. En effet, le stress cellulaire, endommage les protéines, ce qui provoque une détérioration de leurs fonctions. Par exemple, des protéines mal repliées peuvent former des agrégats qui peuvent être à l'origine de différentes maladies neurodégénératives, telles que les maladies de Parkinson et d'Alzheimer (Bosco, LaVoie, Petsko, & Ringe, 2011; Scheper, Nijholt, & Hoozemans, 2011). Afin de lutter contre la formation de ces agrégats, la cellule possède des mécanismes de défense, tels que l'intervention des protéines de choc thermique (Heat shock proteins, HSP), qu'on appelle aussi les protéines « chaperons ». Ces protéines interviennent dans le principal mécanisme lié au repliement des protéines et au contrôle de leur qualité, en formant un réseau qu'on appelle le réseau de l'homéostasie des protéines (RHP). Ce réseau a pour but de contrôler, d'assurer et de protéger le protéome par le biais de la réparation du repliement et l'élimination des agrégats (Fulda, Gorman, Hori, & Samali, 2010; Hartl & Hayer-Hartl, 2002; Milisav, 2011; Richter, Haslbeck, & Buchner, 2010). Ce processus biologique est conservé à travers tous les organismes. Cependant, on remarque qu'en fonction de leurs environnements, différentes espèces possèdent des phénotypes particuliers qui pourraient être reliés à des caractéristiques uniques en RHP. Par exemple, *Northobranchius furzeri*, le poisson killi turqoise présente une perte de l'homéostasie ce qui le rend davantage sensible au stress, ce qui lui confère une longévité plutôt courte (4 à 9 mois) (Y. Kim, Nam, & Valenzano, 2016; Platzer & Englert,

2016). À l'inverse, le rat-taupe nu, alias *Heterocephalus glaber*, possède une incroyable résistance au stress grâce à une bonne régulation de son homéostasie, ce qui lui permet de vivre relativement très longtemps (environ 30 ans), lorsqu'on le compare à la souris (avec une durée de vie moyenne de 4 ans)(Kimberly Riskas, 2018; Musi & Hornsby, 2015). Ainsi, il est nécessaire de se demander, comment l'homéostasie cellulaire fonctionne chez ces deux organismes ainsi que chez d'autres espèces eucaryotes. Cette étude permettrait ainsi de comprendre les principes fondamentaux de l'homéostasie cellulaire par l'étude comparative d'organismes ayant des systèmes de RHP diamétralement opposés.

1.2 Hypothèses et objectifs

Nous émettons l'hypothèse que les espèces qui sont phylogénétiquement proches partagent des caractéristiques communes en chaperons et en composition du protéome[3].

Afin de pouvoir affirmer ou infirmer cette hypothèse, ce projet de recherche consiste à réaliser une analyse comparative des organismes eucaryotes par l'étude (i) de leur nombre de protéines chaperons et (ii) et des propriétés physico-chimiques de leurs protéomes à l'échelle évolutive.

Cette recherche apporte des réponses relatives au nombre des protéines de choc thermique (HSP) et à la composition du protéome sur un nombre élevé d'espèces et qui sont complémentaires aux résultats précédemment publiés par Powers et Balch (Powers & Balch,

[3] Propriétés physico-chimique des protéines du protéome.

2013) qui établissent le nombre de protéines chaperons chez différentes espèces des trois domaines de la vie pour comprendre la diversité de l'origine du RHP.

1.3 Organisation du mémoire

Ce mémoire de maîtrise s'articule en quatre parties. Le présent chapitre présente une courte introduction, une hypothèse ainsi que les objectifs de recherche de ce projet. Dans le chapitre suivant, on fera une synthèse critique des publications antérieures en lien avec le sujet de ce mémoire, d'un point de vue biologique (section 2.1) et bio-informatique (section 2.2). Le troisième chapitre contient l'article scientifique. Finalement, le dernier chapitre présentera une discussion générale qui se conclura d'une perspective.

Chapitre 2

Revue de la littérature

2.1 Aspect biologique

Dans cette section, nous introduirons les notions de base de l'homéostasie cellulaire, puis poursuivrons avec celles relatives au RHP. Plus particulièrement, nous présenterons les chaperons moléculaires ainsi que leur évolution chez les eucaryotes, *Heterocephalus glaber* et *Nothobranchius furzeri*.

2.1.1 Homéostasie cellulaire

L'homéostasie

Le terme « homéostasie » a été introduit par le physiologiste Walter Cannon en 1932. Il le définit comme un état constant des états physico-chimiques simples qui se trouvent dans des systèmes fermés, où les forces connues sont équilibrées. Ces états stables sont maintenus par des processus physiologiques, selon les principes fondamentaux de la régulation et du contrôle (Cannon, 1932).

Au niveau cellulaire, l'homéostasie protéique dite « protéostasie » caractérise l'équilibre du protéome. Cet équilibre se fait par le biais du contrôle de la conformation, de la localisation et de l'activité de chaque protéine. Ce contrôle est réalisé grâce à un réseau de protéines dites « chaperons » qui assure la qualité des protéines par le repliement des protéines nouvellement synthétisées ainsi que par le maintien et la protection de leurs conformations. Les chaperons fonctionnent en conjonction avec les mécanismes de transport et de dégradation des protéines

afin d'assurer l'intégrité du protéome, ainsi que son maintien fonctionnel pour conserver une bonne protéostasie (Y. E. Kim, Hipp, Bracher, Hayer-Hartl, & Ulrich Hartl, 2013).

La défaillance de l'homéostasie

Le stress cellulaire est induit par plusieurs facteurs de stress environnementaux, notamment l'exposition à des toxines (dérivés réactifs de l'oxygène), des températures extrêmes et des dommages mécaniques. Le stress cause le déséquilibre de la protéostasie, et ce, par la défaillance du contrôle de qualité des protéines du RHP induisant ainsi un mauvais repliement des protéines. L'accumulation de ces dernières engendre la formation d'agrégats qui constituent des fibres d'une conformation de feuillet β, appelée « amyloïde ». Les amyloïdes ont de graves conséquences sur le fonctionnement de la cellule et de l'homéostasie cellulaire appelé « protéotoxicité » qui génère des pathologies comme les maladies neurodégénératives ainsi que l'amylose et les troubles métaboliques (Y. E. Kim et al., 2013; Koga, Kaushik, & Cuervo, 2011; Ross & Poirier, 2004).

2.1.2 Réseau de l'homéostasie des protéines

2.1.2.1 Rôles et voies impliquées dans la cellule

Au cours de la vie cellulaire des protéines, certaines dépendent des chaperons pour leurs repliements afin de maintenir leurs conformations fonctionnelles actives. Le RHP joue un rôle clé dans l'évolution continue de la variabilité de la séquence protéique en intervenant sur les propriétés biophysiques du repliement. Ceci permet d'assurer une adaptation optimale à l'environnement et de faciliter une diversité biologique. En outre, le réseau a co-évolué avec le protéome dans le but de favoriser la survie de l'organisme (Powers & Balch, 2013). Lors d'un stress, les chaperons s'expriment afin de faciliter le repliement des protéines mal repliées, et

ce, grâce au facteur du choc thermique (HSF1). Ce facteur entre en compétition avec les protéines non natives qui se forment en conditions du stress. En condition normale, HSF1 forme des homotrimères[4] qui s'associent à la machinerie HSP90/HSP70 et induisent la transcription des gènes cibles (Sala, Bott, & Morimoto, 2017).

Chez les eucaryotes, la cellule se compose d'un système endomembranaire qui possèdent différentes membranes et se divise en compartiments fonctionnels (mitochondrie, lysosome, chloroplaste, organites endocytose et exocytose) (Powers & Balch, 2013). Ces compartiments créent des environnements où se trouvent de nombreuses protéines spécialisées, dont celles pour la gestion de repliement (voir tableau 2.1). Dans ces compartiments, des voies biologiques interconnectées répondent au stress cellulaire en régulant le RHP : la réponse au choc thermique cytosolique (HSR), la réponse aux protéines non pliées (UPR) liée au stress du réticulum endoplasmique (RE) et UPR liée à la mitochondrie, voir leurs implications sur la figure 2.1 (Y. E. Kim et al., 2013; Powers & Balch, 2013).

Les voies de dégradation des protéines non fonctionnelles, les protéines agrégées et leur élimination constituent un élément important du RHP. Ces protéines, qui ne peuvent pas être dépliées par la dégradation du protéasome, peuvent être éliminées par l'autophagie et par la dégradation vacuolaire/lysosomale même en absence de stress. La perte d'autophagie entraîne la formation des agrégats protéique et la neurodégénérescence (Y. E. Kim et al., 2013).

[4] Une protéine composée de trois unités identiques de polypeptide.

Nombre de protéines impliqués	Rôle	Voie impliquée
~ 300	Maintenance de la conformation (repliement, remodelage et désagrégation)	UPR, HSR
~ 400	Biogénèse (transcription, traduction, transport, initiation au repliement)	UPR
~ 700	Dégradation (autophagie, la voie de dégradation ubiquitine)	HSR. UPR

Tableau 2.1 : Protéines impliquées dans les voies biologiques. Tiré de(Y. E. Kim et al., 2013).

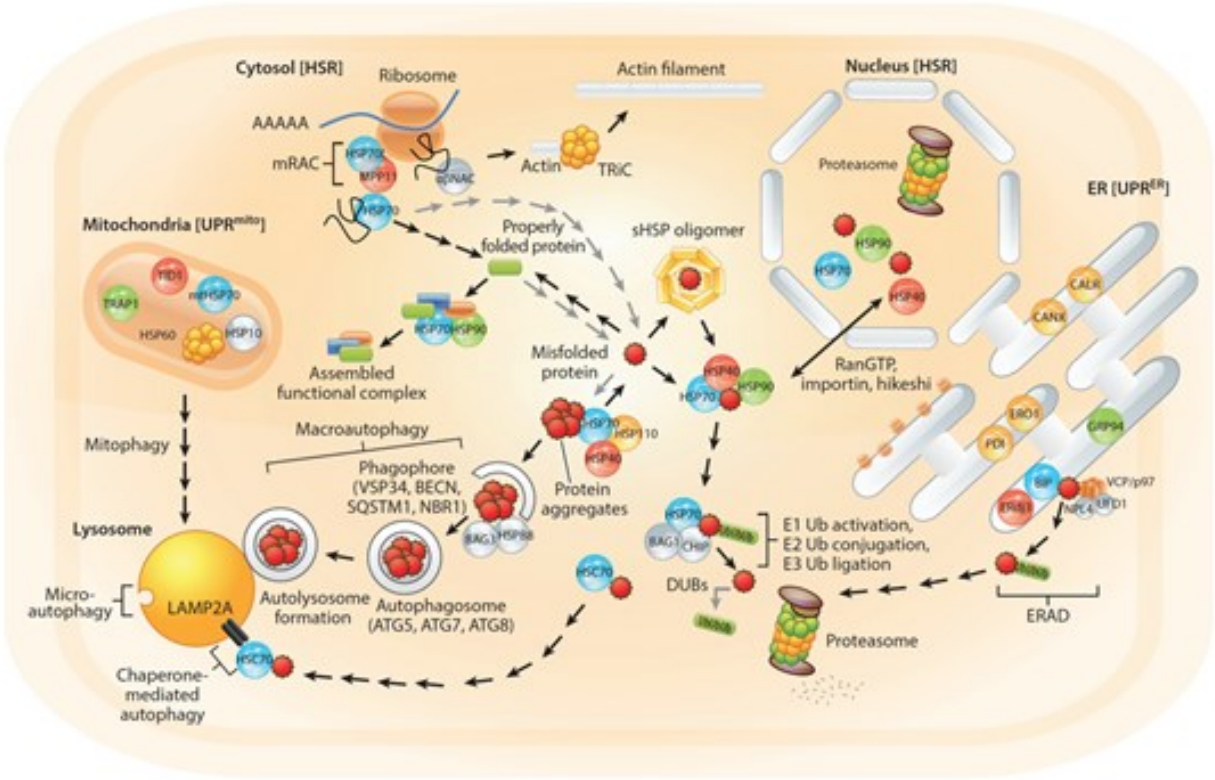

Figure 2.1 : vue d'ensemble du RHP. Tiré de (Labbadia & Morimoto, 2015).

2.1.3 Classes des chaperons

L'origine du terme chaperon moléculaire a été suggérée pour la première fois par Ron Laskey et ses collègues en 1978, suite à la découverte d'une protéine purifiée inconnue liée à l'histone qui avait le rôle de prévention des mauvaises interactions ioniques entre les histones et l'ADN. En 1987, John Ellis définit le terme chaperon moléculaire par « une classe de protéines cellulaire qui ont la fonction d'assurer le pliage de certaines chaines polypeptides et les assembler correctement aux structures oligomériques » (Ellis, 1996).

Plusieurs des chaperons ont été classés comme protéines de stress car elles sont induites par les facteurs de stress à court terme lors d'une réponse d'une haute température ou d'un autre type du stress d'où l'appellation d'HSP. Les chaperons moléculaires se sont très bien conservés chez les organismes vivants tout au long de l'évolution. Ils sont classées d'après leurs poids moléculaires en cinq classes : HSPs connues comme petites protéines de choc thermique dont leur poids est entre 12 et 43 KDa (dont HSP20 et HSP40), HSP60 - 60 KDa, HSP70 - 70 KDa, HSP90 - 90 KDa et HSP100 - 100 KDa (Koga et al., 2011).

HSP20

La classe d'HSP20 est la plus diversifiée en nombre, d'un seul type chez *Saccharomyces cereviceae* à une trentaine chez les plantes, et en poids moléculaire, de 16 KDa chez *Caenorhabditis elegens* à 40 KDa chez *Schistosoma mansoni*. Malgré cette diversité, les petits chaperons ont une grande homologie au sein d'une même espèce: 90 % d'identité entre les organismes du même sous-groupe (cas de *Glycine*) et 20 % d'identité entre les organismes éloignés (cas du *Caenorhabditis elegans*, *Drosophila melanogaster* et *Xénope lisse*). L'homologie est similaire sur le profil d'hydrophobicité et sur quelques régions protéiques (Lindquist & Craig, 1988). HSP20 est conservé par sa structure ainsi que par sa séquence

protéique (voir figure 2.2). En effet, au niveau C-terminale, la protéine contient un domaine d'alpha cristalline d'environ 100 acides aminés (AA) qui est présent chez tous les HSPs et qui aide à l'activité de chaperon. Au niveau du N-terminale, la protéine apparentée à la protéine Ras R-Ras (RRAS) contient un motif consensus peptidique, qui est phosphorylé par deux protéines kinases A et G (PKA et PKG) à la sérine. Cette phosphorylation, qui peut jouer un rôle au niveau des propriétés anti-apoptotiques, agit comme un médiateur de cardio-protection. Par exemple, des études menées par Dr. Evangelia Kranias révèlent que HSP20 a la capacité à protéger le cœur contre les lésions induites par la reperfusion et par l'ischémie[5] (Edwards, Scott, & Baillie, 2012; Evangella Kranias, 2015; Nicolaou et al., 2008).

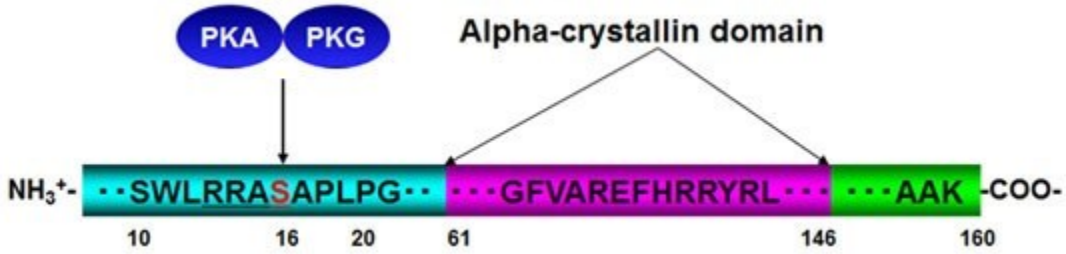

Figure 2.2 : Séquence primaire d'HSP20 chez l'humain. Tiré de (Evangella Kranias, 2015).

HSP40, HSP70

HSP70 est très conservé chez les eucaryotes et les procaryotes. En condition normale et de stress à court terme, HSP70 a un rôle essentiel dans le fonctionnement des protéines. Dans un cycle de réaction hydrolytique d'ATP, le chaperon moléculaire se lie et libère des régions hydrophobes d'une chaîne polypeptidique. Le cycle de liaison et de libération est également modulé par HSP70 et HSP40. La classe d'HSP40 est reconnue par sa conservation du domaine J. Ce domaine est responsable de l'interaction entre HSP40 et HSP70, qui se fait par une

[5] L'arrêt ou l'insuffisance de la circulation sanguine.

liaison polypeptide. Ce système HSP70/HSP40 a pour rôle de protéger les protéines sujettes à l'agrégation (Hartl, 1996).

La protéine HSP40 est similaire à la protéine chaperone DnaJ chez les bactéries, car elle contient plusieurs segments communs. Sa structure (voir figure 2.3), au niveau N-terminal, consiste en un domaine J de 70 AA suivi d'une région riche en glycines et phénylalanines d'une longueur de 30 AA. Ces deux surfaces constituent la région d'interaction d'HSP70. Le domaine suivant de 80 AA contient deux atomes de zinc et, au niveau C-terminal, se trouve une région de 165 AA. Ces deux derniers segments permettent à HSP40 de se lier au substrat et d'interagir avec HSP70 (Hartl, 1996).

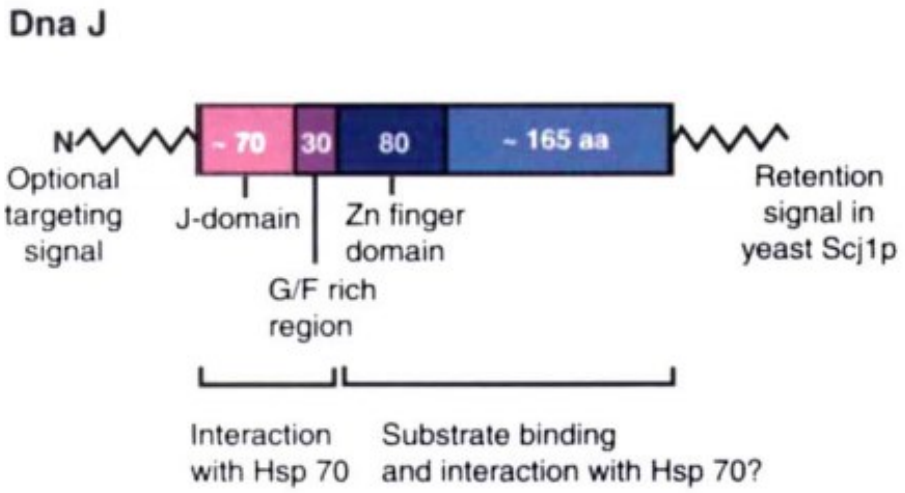

Figure 2.3 : Structure primaire de HSP40. Tiré de (Hartl, 1996).

Quant à la structure de HSP70 (voir figure 2.4), on trouve, au niveau N-terminal, le domaine d'ATPase (environ 385 AA), suivi d'un domaine de liaison de peptides (environ 225 AA) et, au niveau C-terminal, 30 AA riches en proline et glycine qui contient un motif Glu–Glu–Val–Asp (EEVD) (environ 30 AA) (Hartl, 1996).

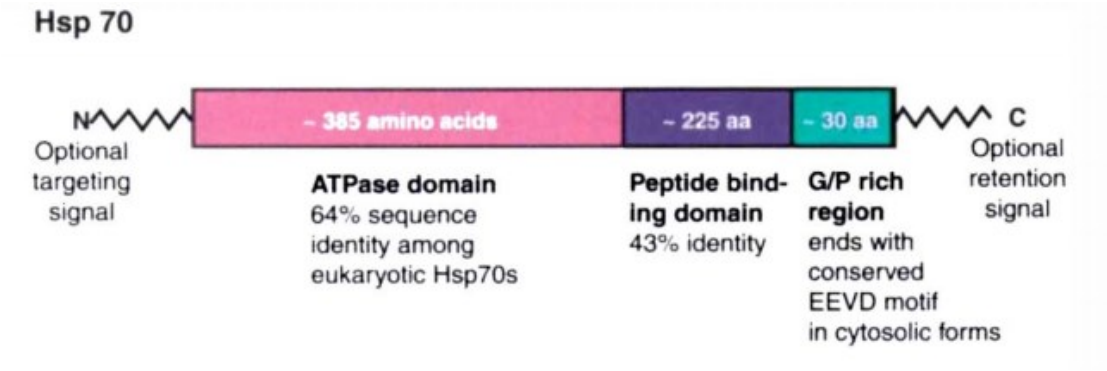

Figure 2.4 : Structure primaire de HSP70. Tiré de (Hartl, 1996).

HSP60

Aussi appelée chaperonine, HSP60 joue un rôle dans le repliement des protéines, en condition normale et de stress à court terme et dépends de l'ATP. Cette protéine est divisée en deux sous-groupes en raison de leurs différentes structures et de quelques spécificités sur leurs fonctions. Le groupe 1 de type de GroEL est présent chez les eubactéries (voir sa structure primaire sur la figure 2.5) alors que le groupe 2 TriC (tailles complexes polypeptide-1 (TCP-1) du complexe d'anneau) ou CTT (chaperonine contenant TCP-1) est présent chez les archées et les eucaryotes (Hartl, 1996).

Nous allons particulièrement nous intéresser au second sous-groupe étant donné que notre étude se centre surtout sur les organismes eucaryotes.

TriC a un rôle plus spécifique dans le repliement ; il replie l'actine et la tubuline, mais aussi plusieurs sous-ensembles de polypeptides en in-vivo. Il a pu co-évoluer avec des protéines eucaryotiques afin de stabiliser la productivité du repliement intermédiaire, car lui seul a la capacité de produire des repliements intermédiaires aux protéines pour les faire passer à leurs états natifs (Hartl, 1996). Au niveau de sa structure, la molécule est formée d'anneaux double de 800 à 1000 Kda avec une cavité centrale qui permet d'améliorer le pliage de protéines que

le système HSP70 n'est pas capable de plier. La structure de TRiC est très semblable à GroEL. Cette dernière a aussi une structure à anneaux doubles, où chaque anneau est constitué de 8 sous-unités (Frydman, 2001; Hartl, 1996; Y. E. Kim et al., 2013).

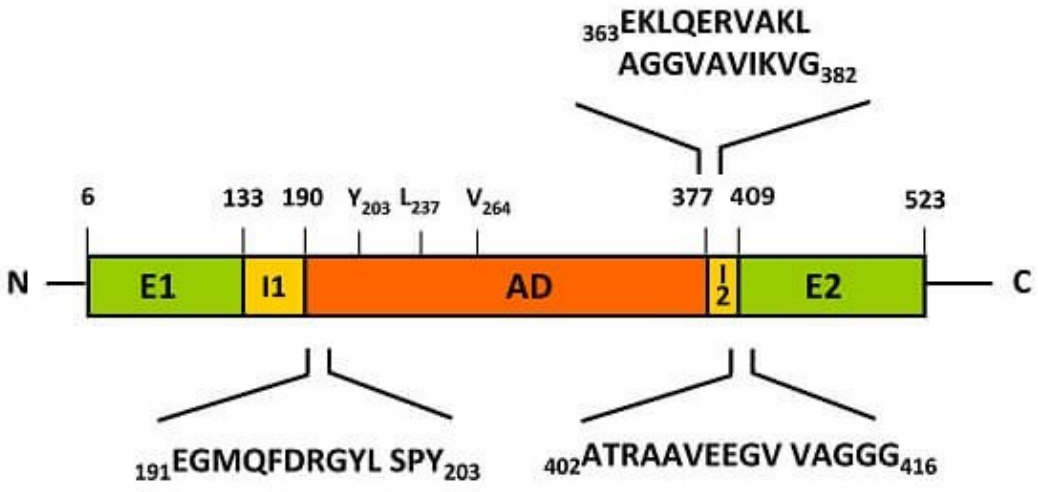

Figure 2.5 : Structure primaire d'HSP60 GroEL chez les procaryotes. Tiré (Jürgen Radons, 2018a).

HSP90

Ce chaperon moléculaire est hautement conservé et abondant aux températures normales et élevées. Au niveau C-terminal, la protéine contient une région conservée (MEEVD) qui a le rôle de liaison avec le co-chaperon au domaine de répétition tétratricopeptide (TPR). Au milieu de la séquence, se trouve le site d'ATPase et de liaison des protéines cibles. Enfin, au niveau N-terminal il y a le domaine de liaison et l'hydrolyse d'ATP (figure 2.6). HSP90 joue un rôle essentiel dans l'interaction avec les récepteurs d'hormones stéroïdiens qui participent au mécanisme de transduction du signal (Picard et al., 1990). En effet, les modifications post-traductionnelle comme la S-nitrosylation, la phosphorylation et l'acétylation, contrôlent l'activité HSP90. Ces modifications influencent également l'interaction entre HSP90 et les co-

chaperons en affectant le pliage des protéines clients, ce qui provoque un stress protéotoxique (Ali et al., 2006; Jürgen Radons, 2018c; Leach, Klipp, Cowen, & Brown, 2012).

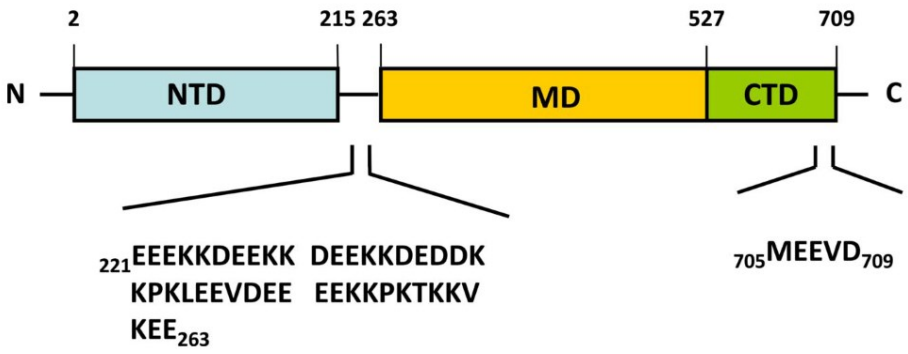

Figure 2.6 : Structure primaire d'HSP90 du champignon. Tiré de (Jürgen Radons, 2018c).

HSP100/Clp (Protéase caséinolytique P)

Ce chaperon appartient à la super famille AAA (ATPases Associées à diverses Activités cellulaires). Il collabore avec le système HSP70 afin de déplier les protéines mal repliées ou des protéines marquées. HSP100 a deux domaines distincts, un domaine α-hélicoïdal et un domaine de liaison aux nucléotides de type Walker. Il est divisé en deux classes (figure 2.7) : la classe I regroupe les protéines aux modules AAA et la classe II, les chaperons. Les protéines de classe I partagent deux domaines de liaison nucléotides (NBD-1 et NBD-2) et un domaine milieu qui peut être absent chez quelques protéines. De leur côté, les protéines de classe II ne partagent que le second domaine de liaison nucléotides (NBD-2). Le chaperon est analogue au protéosome par son complexe ClpA (chaperon) et ClpP (protéase) et est souvent repéré en association avec un anneau de protéase. En effet, si une protéine est marquée, ClpA la déplie avec une activité ATP et ClpP la dégrade. En cas d'absence de ClpP, la protéine est relâchée et se replie à son état d'équilibre (Patrick D'Silva, 2014; Saibil, 2000).

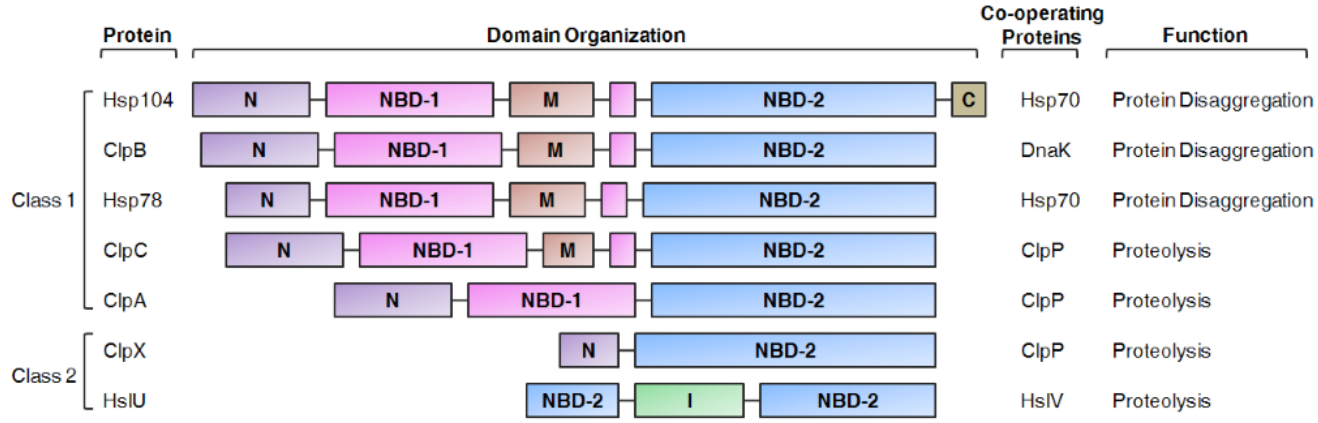

Figure 2.7 : Structure primaire d'HSP100/Clp. Tiré de (Patrick D'Silva, 2014).

Nous résumons les fonctions des chaperons sous forme d'un tableau :

Fonctions chez les eucaryotes

HSP 20	Maintien du repliement des protéines grâce une activité de chaperon, prévention de la formation d'agrégats, ce qui lui donne sa fonction anti-apoptotique, consolidation du cytosquelette lors d'un stress ischémique, stabili du génome, maintien de l'intégrité myofibrillaire (Bakthisaran, Tangirala, & Rao, 2015; Edwards et al., 2012)
HSP 40	Responsable de la régulation de l'activité ATPase par sa liaison avec HSP70, liaison de chaines naissantes, prévention d la formation des protéines agrégats (Greene, Maskos, & Landry, 1998).
HSP 60	Favorise le repliement des protéines et leur assemblage, présente une activité ATPase intrinsèque, prévention de la formation des protéines agrégats(Jürgen Radons, 2018b).
HSP 70	Repliement, assistance, contrôle, dégradation et transport des protéines, système contrôle de qualité des protéines, maintien de l'homéostasie des protéines durant le stress (Uniprot, 2018).
HSP 90	Assistance des protéines, maintien de la structure tertiaire du protéasome pour la dégradation des protéines mal repliées, régulation de l'hormone stéroïde par son interaction avec le récepteur des glucocorticoïdes (Imai, Maruya, Yashiroda, Yahara, & Tanaka, 2003; Pratt, Morishima, Murphy, & Harrell, 2006).
HSP 100	Désagrégation et dépliement des protéines mal repliées ou marquées, protéolyse (Saibil, 2000).

Tableau 2.2 : Fonctions des HSP chez les eucaryotes.

2.1.4 Évolution du réseau d'homéostasie des protéines

Dans cette section, nous allons détailler l'évolution du RHP relié à la voie HSR ainsi que la coévolution entre les chaperons et le génome.

a) – Eucaryotes

Évolution du réseau de l'homéostasie des protéines

Le RHP et le protéome ont co-évolué pour la survie des organismes. En effet, le réseau facilite à la fois l'adaptation et la sélection naturelle des organismes dans leur environnement, et ce, par des mécanismes génétiques et épigénétiques. Par exemple, HSP90 peut gérer les mécanismes liés aux transposons et peut agir comme un activateur ou inhibiteur de ce processus par ses effets sur la structure de la chromatine qui gèrent les relations structure-fonction des nucléosomes à base d'histones et de la stabilité des télomères. En effet, HSP90 peut affecter la liaison de chromatine de l'enzyme histone acétyltransférase (HAT), responsable du remodelage et de la décondensation de la chromatine, et de l'enzyme histones désacétylases (HDAC), responsable de la condensation de la chromatine. Ces deux enzymes régulent l'expression de HSP40, HSP70 et HSP90 ainsi que l'activité du facteur de transcription HSF1 qui peut baisser l'expression des chaperons (la réponse HSR) (Powers & Balch, 2013). En raison de cette expression omniprésente chez toutes les espèces eucaryotiques qui contribue en tant qu'un mécanisme de défense indispensable pour la protection des cellules contre les conditions environnementaux, stress oxydatif, métaux lourds, inflammation, fièvre (Jolly & Morimoto, 2000). Plusieurs études montrent une qu'il y a une forte expression des chaperons lors d'un choc thermique, ce qui détermine que les chaperons

jouent un rôle d'adaptation au sein d'espèces ainsi que des biomarqueurs du stress thermique (Sørensen, 2010).

Coévolution des chaperons moléculaires et du génome

Les chaperons sont largement distribués, mais pas uniformément, dans les trois domaines du vivant : Eucaryote, Archée et Bactéries (voir la figure 2.8 et tableau 2.S1 sur annexe) (Powers & Balch, 2013). D'après la figure 2.8, l'expression des chaperons moléculaires à HSR est corrélée avec la taille du génome. Plus spécifiquement, l'augmentation de la complexité des protéines fonctionnelles et structurales a évolué avec la protéostasie (figure 2.9). En moyenne, pour chaque 2000 gènes présents chez un organisme, il y a un homologue d'HSP20, cinq à six homologues d'HSP40, un homologue d'HSP60 et un d'HSP70 et pour chaque 6000 gènes il y a un type d'homologue d'HSP90. La corrélation entre les chaperons moléculaires et la taille du génome n'est pas bien comprise, car les HSPs n'ont pas une forte spécificité au substrat d'où le fait, par exemple, qu'un niveau élevé d'expression d'un seul type d'HSP70 est suffisant pour maintenir la protéostasie (Powers & Balch, 2013).

C'est à partir de cette réflexion que nous avons eu l'idée d'estimer le nombre de chaperons en fonction du nombre des protéines dans le protéome par des outils bio-informatiques sur un nombre élevé d'organisme du règne d'eucaryotes afin de voir l'adaptation et la coévolution du protéome avec les chaperons moléculaires.

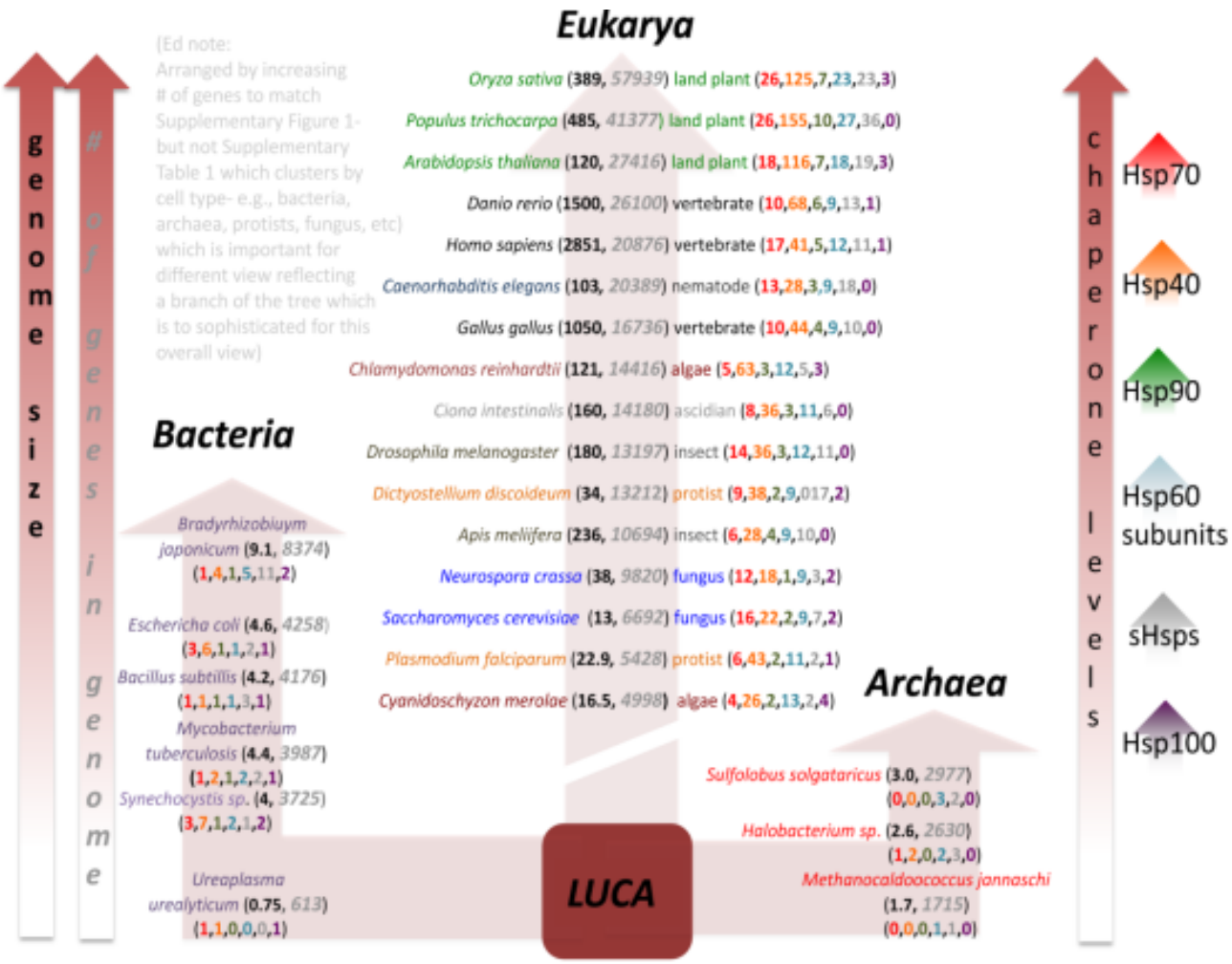

Figure 2.8 : Évolution des chaperons moléculaires chez trois domaines du vivant. Les nombres représentés en figure après le nom de l'espèce sont la taille du génome en noire et le nombre de gènes dans le génome en gris. Quant aux nombres chaperons sont représentés respectivement HSP20 en gris, HSP40 en orange, HSP60 en bleu ciel, HSP70 en rouge, HSP90 en vert et HSP100 en violet. Tiré de(Powers & Balch, 2013).

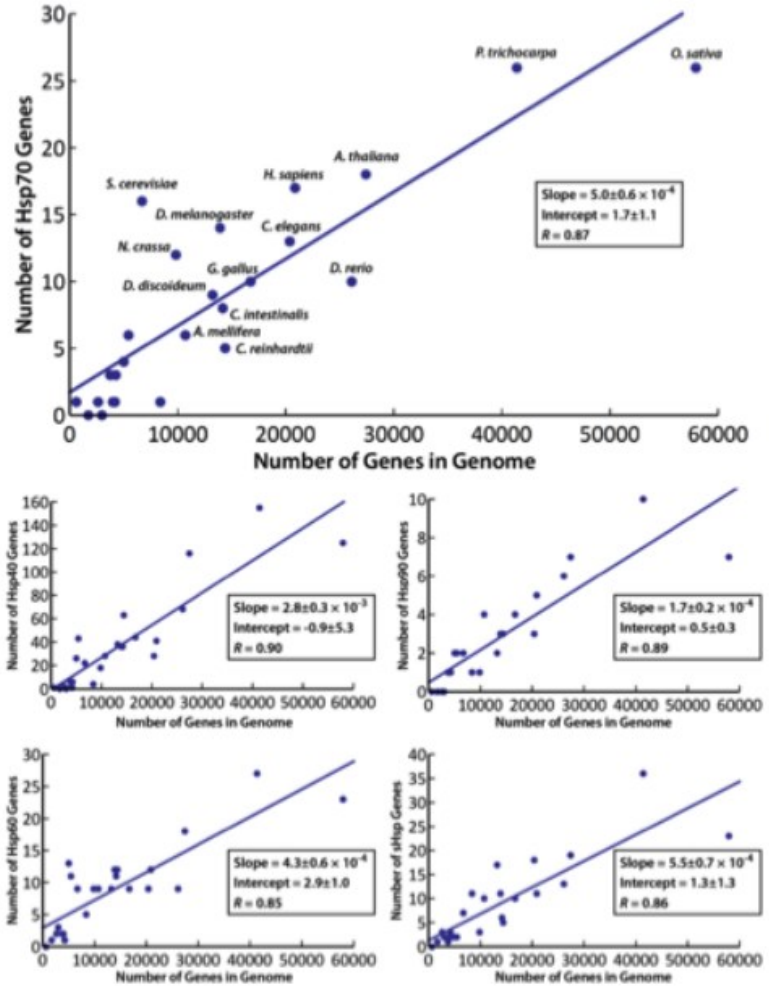

Figure 2.9 : Corrélation entre le nombre de gènes dans le génome et le nombre des chaperons. Tiré de (Powers & Balch, 2013).

b) -Heterocephalus glaber

H.glaber plus communément appelé rat-taupe nu appartient au phylum Chordata (Myers, 2018). Cette espèce vit naturellement dans les régions chaudes et arides du nord-est de l'Afrique où elle mène une existence strictement souterraine, se nourrissant d'organes de stockage de plantes trouvés sous terre (Musi & Hornsby, 2015). *H.glaber* a évolué dans un

environnement très protégé, énergiquement exigeant et durement agressif, ce qui aurait permis de mettre au point des mécanismes assurant la longévité (S. Eddy, 1992). Depuis le début des années 2000, l'importance de *H.glaber* en tant que modèle de longévité unique, avec ses caractéristiques de vieillissement et sa résistance au cancer, en fait une espèce sujette à des recherches dans de nombreux laboratoires (Musi & Hornsby, 2015).

Avec le progrès technologique du séquençage, plusieurs laboratoires se sont penchés sur la nature du génome du *H.glaber* et sur les changements pouvant survenir au cours du vieillissement. Il semble être capable de maintenir l'expression génomique de plusieurs de ses gènes entre l'âge de 4 à 20 ans dans les tissus du cerveau, des reins et du foie. Ces gènes particuliers sont impliqués dans un mécanisme supérieur de réparation de l'ADN qui aurait un rôle à jouer dans la longévité du rongeur (Musi & Hornsby, 2015).

Lors d'un stress, il y a une résistance aux dommages oxydatifs chez *H.glaber*, ce qui semble indiquer une bonne protection du protéome. Ceci s'explique par ses protéines qui résistent aux dommages oxydatifs et à la chaleur, car elles présentent une stabilité structurelle protéique supérieure aux protéines des souris, vu les niveaux plus élevés de dommages protéiques réversibles plutôt qu'irréversibles (Musi & Hornsby, 2015).

Le compartiment clé de l'espèce est le protéasome. En effet, *H.glaber* a une activité très importante du protéasome par rapport à celui de la souris, vu la présence d'immunoprotéasome[6]. Ces immunoprotéasomes ont des sous-unités catalytiques distinctes et des régulateurs qui sont très efficaces dans les activités protéolytiques. Une autre étude

[6] Un autre type du protéasome sous forme de trois sous-unités β catalytiques associé en réponse immunitaire.

comparative du rat-taupe nu avec la souris et l'humain, montre que les protéasomes du *H.glaber* semblent imperméables à l'inhibition par des dommages oxydatifs ou par des inhibiteurs spécifiques du protéasome. Cette résistance de l'activité du protéasome face aux agents inhibiteurs contribue au maintien de l'homéostasie des protéines même en présence de résistance au stress, sans nécessiter d'autophagie pour éliminer les agrégats de protéines plus importants et les organites endommagés (Musi & Hornsby, 2015).

c) – Nothobranchius furzeri

N.furzeri plus communément appelé killi turquoise, appartient au phylum chordata (Myers, 2018). Ce poisson vit dans des mares éphémères au zone semi-aride au Zimbabwe. D'une maturation accélérée, son court cycle de vie est de 4 à 6 mois (Harel, Valenzano, & Brunet, 2016). Parce qu'il possède un cycle de vie très court, il est un modèle par excellence dans la génomique comparative qui cherche à étudier le vieillissement (Terzibasi, Valenzano, & Cellerino, 2007). En effet, les différentes caractéristiques pertinentes à l'étude du vieillissement du *N.furzeri* sont la sénescence cellulaire, l'altération épigénétique et la communication intracellulaires, la perte de protéostasie, l'instabilité génomique en abrasion des télomères ainsi que le dysfonctionnement mitochondriale. *N.furzeri* surexprime les gènes codant pour les protéines lysosomales ainsi que pour la régulation à la hausse pour la traduction, et ce, malgré la perte de la RHP. En effet, l'accumulation d'agrégats lysosomales est une caractéristique du vieillissement neuronal. De plus, *N.furzeri* a une baisse de régulation de protéasome responsable de la suppression des agrégats (Baumgart et al., 2014; Platzer & Englert, 2016).

2.2 Aspect bio-informatique

Afin de mieux comprendre les principes d'évolution du système d'homéostasie des protéines, précisément, les chaperons. Une analyse comparative sera faite à travers les organismes eucaryotes (par des méthodes phylogénétiques) qui requière tout d'abord l'identification des chaperons (par le modèle de Markov caché (HMM)). Dans cette section, nous introduirons quelques concepts en bio-informatique utilisés lors du projet. D'abord, nous allons décrire l'outil basé sur les HMM. Ensuite, nous allons introduire la phylogénie ainsi qu'un aperçu de certaines méthodes et outils bio-informatiques qui seront utilisés pour inférer l'évolution du RHP.

2.2.1 Identification des chaperons

Afin de classer les séquences inconnues au sein des familles protéiques d'HSP, deux étapes ont été faite. La première est la collecte des profils d'HMM sur une base de données. La seconde est la recherche des HSP sur différentes classes à partir des profils d'HMM par le progiciel HMMER (S. Eddy, 1992).

Les familles protéiques contiennent des régions conservées qui ont souvent reliés à une fonction enzymatique importante qui distingue une famille à une autre. Ces régions se réfèrent comme des domaines de familles (Horan, Shelton, & Girke, 2010). Ces familles peuvent aussi contenir des régions peu conservés (insertions/délétions), identifiées par des alignements. Les méthodes de profil HMM permettent de gérer de façon assez robuste la conservation des séquences et les insertions/délétions (Sonnhammer, Eddy, Birney, Bateman, & Durbin, 1998). Il est possible de générer un profil d'HMM à partir d'un alignement multiple, qui peut être utilisé pour une recherche des familles de protéines dans une base de données. Les régions

conservées dans l'alignement représentent les principaux états du modèle HMM. De leur côté, les régions non conservées représentent des états d'insertions, car l'alignement est incertain. Le but est donc de définir les probabilités aux fréquences d'acides aminés (AA) pour chaque état principal observé (par la multiplication de toutes les probabilités) et d'estimer les probabilités de transition d'un AA à un autre à partir des distributions de probabilités des acides aminés. La recherche sur la base de données se fait par l'alignement sur le modèle HMM. Ce dernier consiste en une affectation d'état à chaque résidu de la séquence qui génère une grande possibilité d'alignements. Or, il existe un algorithme efficace « l'algorithme Viterbi » qui utilise la programmation dynamique et donne le score et la probabilité des séquences d'alignement (Anders Krogh, 1998). Plus précisément, l'algorithme cherche le chemin le plus probable à travers l'HMM pour chaque séquence, et ce, par le calcul de la probabilité du chemin maximum (Choo, Tong, & Zhang, 2004). Les profils d'HMM sont implémentés dans plusieurs programme comme HMMER (Choo et al., 2004).

Pour ce projet, les profils d'HMM ont été récoltés à partir de la base de données d'HSP (HSPIR) conçu par le laboratoire du Dr Patrick D'Silva (Sinha, Veedin Rajan, Esthaki, & D'Silva, 2012) . Celle-ci contient environ 9900 protéines de 277 génomes englobant les procaryotes et les eucaryotes (Sinha et al., 2012).

Les modèles d'HMM sont les meilleurs outils comparant à BLAST et PSI-BLAST pour l'identification des domaines protéiques au sein d'une famille (Söding, 2005). L'outil HMMER applique une recherche des séquences inconnues par une classification en sous-famille qui se fait par la structure et en superfamille qui se fait par diverses fonctions. La recherche procède en deux étapes : trouver les séquences contenant les domaines conservés de la superfamille, puis, chercher la signature de la sous-famille. Si ces deux étapes trouvent le

domaine conservé et la signature, cela veut dire que la séquence appartient à la sous-famille qui a la valeur de E la plus basse. Autrement, la séquence est classée comme une nouvelle superfamille de protéines (Choo et al., 2004).

2.2.2 Composition du protéome

Les chaperons moléculaires sont principalement responsables du repliement des protéines, ils les maintiennent et protègent contre l'agrégation. En effet, les protéines d'une grande taille sont plus difficiles à être repliés en comparaison avec les protéines de taille moyenne. Les forces hydrophobes sont les principaux déterminant du repliement ou l'agrégation des protéines (S. Pechmann, Levy, Tartaglia, & Vendruscolo, 2009). En effet, l'hydrophobicité donnent une stabilité à la protéine ou en la conduisant à un changement irréversible à une conformation collante riche en feuilles bêta qui mènent à l'agrégation (Reynaud, 2010). Cependant, il existe une forte pression évolutive que la cellule exerce pour qu'elle évite l'agrégation de protéines car cette dernière soumet l'RHP à une contrainte extrême qui peut engendrer la mort cellulaire et générer des maladies neurodégénératives (S. Pechmann et al., 2009). Le mauvais repliement ou l'agrégation peuvent être approximé à partir d'une séquence protéique et doivent en principe nous indiquer le nombre de complexité des protéines qu'un organisme a créé. Ceci nous amène donc à étudier les propriétés physicochimiques des protéines, afin de comprendre les mécanismes utilisés par la cellule pour favoriser sa survie en évitant l'agrégation des protéines et en améliorant leurs solubilités. C'est donc pour ces raisons que nous avons choisi d'étudier la propriété physicochimique qu'est l'agrégation.

Des algorithmes spécialisés ont été créés pour mieux comprendre l'agrégation et la vue d'ensemble du RHP. Parmi eux, nous citons : Zyggregator (Gian Gaetano Tartaglia &

Vendruscolo, 2008) et Tango (Fernandez-Escamilla, Rousseau, Schymkowitz, & Serrano, 2004). L'algorithme de Tango prédit la propension d'agrégation des peptides, ou structure bêta croisées, qui mène à la formation de fibres amyloïdes. Cette dernière est calculée en fonction de quatre contributions énergiques :(i) coût en entropie pour chaque AA; (ii) interaction de chaque AA en ajoutant deux paramètres de structure et de position; (iii) interaction de chaine principale ou de chaine latérale; (iv) interaction entre les chaines principales i à i+3 qui se trouvent une liaison hydrogène. Ce programme émet un fichier de sortie qui donne le pourcentage d'agrégation pour chaque AA. En règle générale, sur 5 à 6 des résidus, si l'agrégation est supérieure à 5%, cette région présente un segment d'agrégation potentiel (Luis Serrano ; Joost Schymkowitz ; Frederic Rousseau, 2018).

2.2.3 Méthode phylogénie

3.2.3.1 Phylogénie

C'est dans l'ouvrage « *On the origin of species* », publié en 1859, que Charles Darwin élabore sa théorie de l'évolution. D'après cette théorie, c'est grâce au processus de sélection naturelle et à l'accumulation des différences entre les espèces, que ces dernières évoluent au fil des générations. Ces processus expliquent également la diversité biologique et permettent de représenter la filiation de parenté entre les espèces sous la forme d'un arbre. Historiquement, le terme « phylogénie » fut initié par le biologiste allemand Ernst Haeckel, en 1866, pour désigner le lien de filiation qui associe les êtes vivants. Haeckel avait pour but de créer un arbre généalogique qui représente « le système naturel », idée issue de la théorie de Darwin. Il créa un arbre monophylétique[7], composé de trois royaumes (Plantae, Protista, Animalia) (voir figure 2.10). L'arbre illustre la série des étapes morphologiques de l'histoire évolutive d'une espèce, et non la séquence de ses ancêtres. Haeckel a également initié l'utilisation de nouveaux termes en phylogénie (Darlu & Tassy, 1993; Dayrat, 2003; futura planète, 2018; M. Mariadassou ; A. Bar-Hen, 2010).

C'est au XXe siècle, après la découverte de l'existence de différents types de cellules et l'avancement génétique, qu'il y a eu lieu la découverte de l'origine des variations génétiques et de l'héritabilité des espèces, confirmant donc le mécanisme darwinien (Damien Aubert, 2017). En effet, les travaux de Woese ont pu diviser le domaine de la vie en trois : (*i*) eubactérie ; (*ii*) archaebactérie et (iii) eucaryotes, en se basant sur l'analyse d'ARNr 18S pour

[7] Groupe d'organismes qui forme un clade.

les eucaryotes (plantes , animaux et champignons) et d'ARNr 16S/18S pour les procaryotes (Woese & Fox, 1977). Désormais, les noms des trois domaines sont Bactéries, Eucaryotes et Archées(Woese, Kandler, & Wheelis, 1990).

Grâce à la disponibilité des données génomiques, une nouvelle représentation de l'arbre de la vie (Hug et al., 2016) (voir figure 2.11) basée sur la concaténation d'un ensemble de protéines ribosomales, a mis en évidence une forte distribution des bactéries dans l'arbre. L'emplacement des eucaryotes par rapport aux Bactéries et Archées est controversé et n'est toujours pas résolu selon la dernière représentation de l'arbre (Hug et al., 2016).

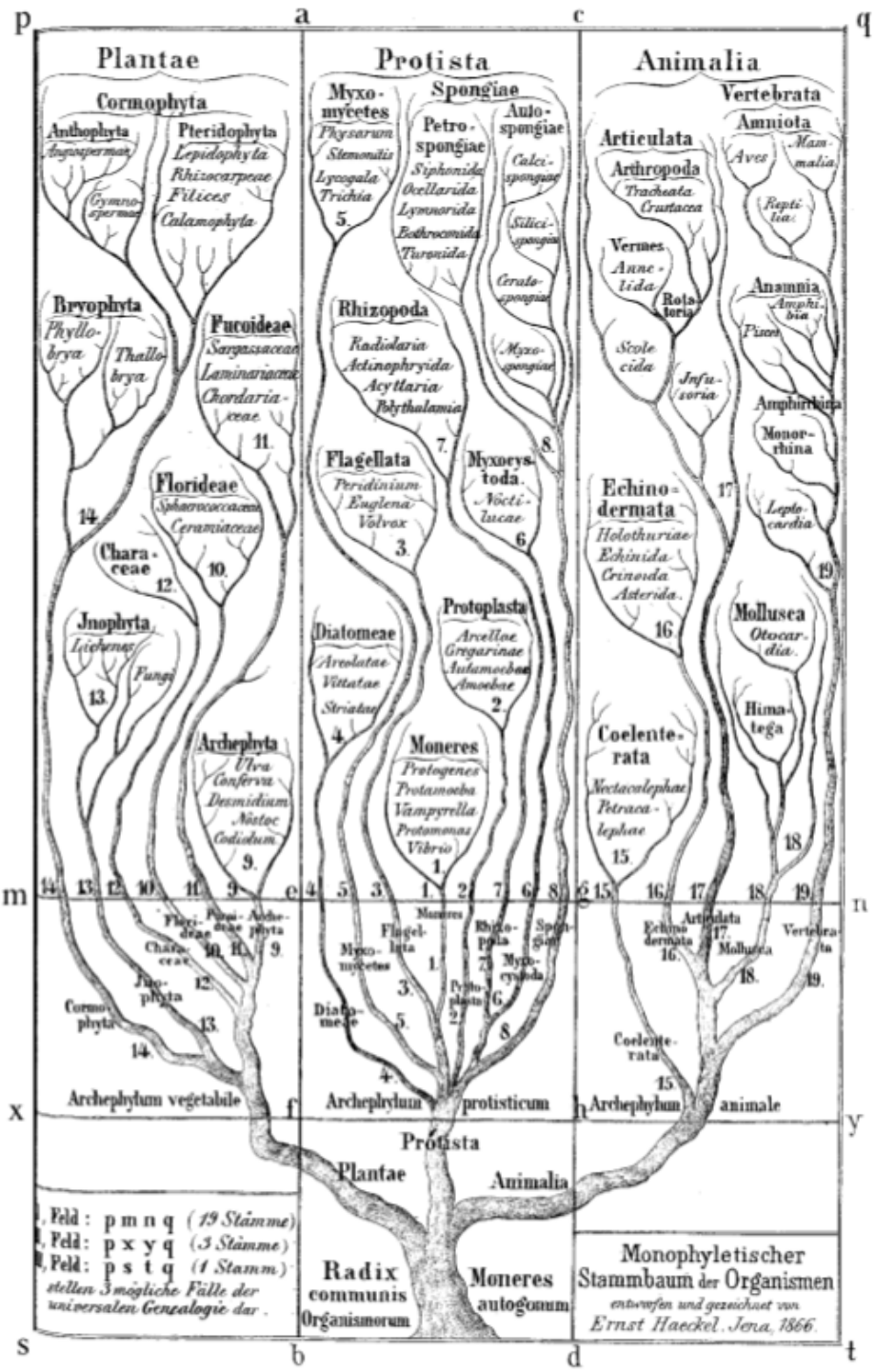

Figure 2.10 : Première représentation de l'arbre de la vie. Tiré de (Dayrat, 2003).

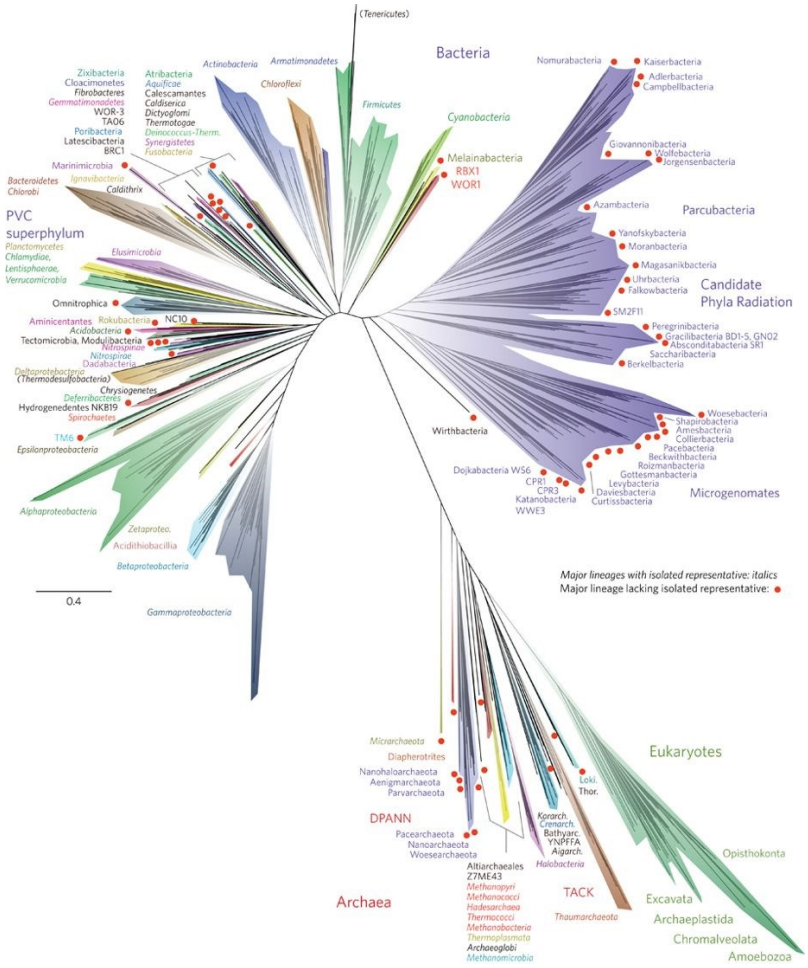

Figure 2.11 : Nouvelle représentation de l'arbre de la vie. Tiré de (Hug et al., 2016).

3.2.3.2 De la phylogénie à la phylogénomique

La phylogénomique a été utilisée pour la première fois dans la prédiction des fonctions des gènes dans le contexte évolutif (Eisen, 1998). La phylogénomique est devenue un outil indispensable et un nouveau domaine en phylogénie (Burki, 2014). On dispose aujourd'hui d'énorme données génomiques. Par conséquent, l'analyse de ces données requiert plus de traitement, ce qui a favorisé une croissance et une diversification rapides des méthodes, des moyens et des algorithmes de calcul (Béatrice Roure, 2011). Ces moyens promeuvent des

nouvelles méthodes d'inférence en phylogénomique, comme l'utilisation d'un simple marqueur ou marqueurs multiple (Wu & Eisen, 2008). Entre outre, la phylogénomique applique deux approches : (i) super-arbre, soit une combinaison d'un ensemble d'arbres provenant d'une analyse de gènes séparément (de Queiroz & Gatesy, 2007) ; (ii) super-matrice, soit la concaténation de plusieurs données (marqueurs) en une seule matrice phylogénétique qui est ensuite utilisée pour faire des inférences (de Queiroz & Gatesy, 2007).

Les nouvelles méthodes d'inférence phylogénomique par plusieurs marqueurs protéiques présentent cependant certaines limites. En effet, au niveau de l'étape d'alignement, il peut par exemple y avoir un manque de données entre les marqueurs ce qui donne des blocs vides sur l'alignement. Malgré l'existence de l'étape de masquage qui consiste à supprimer les positions qui contiennent des espaces vides (gaps) ou à faible degré de conservation, ceci peut néanmoins favoriser la suppression des sites informatifs. Par conséquent, cela cause un manque de résolution ou des artéfacts sur l'arbre. Ces nouvelles méthodes sont également limitées au niveau de leur représentation des lignées d'espèces de l'arbre des gènes puisque les copies d'un gène peuvent ne pas correspondre à l'arbre des espèces. Ce genre d'erreur peut être causé par un polymorphisme génétique, une duplication des gènes, une hybridation ou un transfert horizontal de gènes (de Queiroz & Gatesy, 2007; Wu & Eisen, 2008).

Le marqueur ARNr 18S est suffisant pour créer un arbre à l'échelle des eucaryotes. En effet, il est préférable de choisir un seul marqueur qui évolue lentement pour les groupes divergents (Patwardhan, Ray, & Roy, 2014). En effet, dans le cas contraire, les marqueurs d'ADN mitochondriaux non codants sont utilisés pour l'étude des individus d'une seule population

(Patwardhan et al., 2014). Dans la prochaine section, nous donnerons davantage de détails sur le marqueur choisi et son importance en phylogénie des eucaryotes.

3.2.3.3 Marqueur moléculaire

Séquence ARNr 18S

Le ribosome est composé de deux unités majeures, soit les protéines ribosomales et l'ARNr. La petite sous-unité ribosomale contient un seul type d'ARNr et de 30 protéines ribosomales chez les eucaryotes. La large sous-unité ribosomale contient 3 ARNr et 40 protéines ribosomales. L'ADN ribosomal (ADNr) (figure 2.12) se compose d'un segment unitaire en répétition en tandem, d'un espaceur non transcrit (NTS), d'un espaceur transcrit externe (ETS), d'un espaceur non transcrit 1 et 2 (ITS1, ITS2) et enfin des séquences 5.8S (environ 160 nucléotides), 18S (environ 1800 nucléotides) et 28S (plus que 4000 nucléotides) (Hillis & Dixon, 1991).

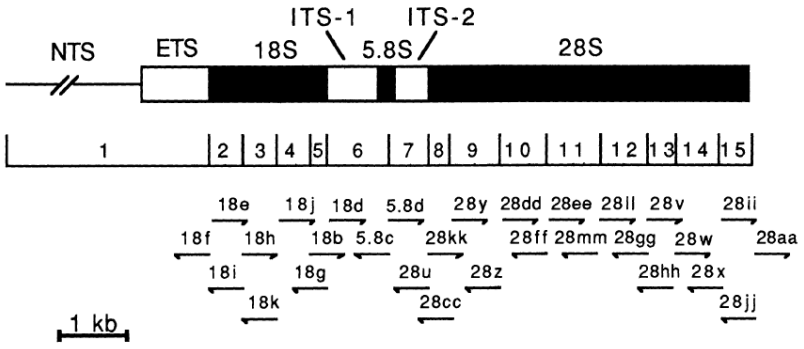

Figure 2.12 : Structure primaire d'ADN ribosomal chez les eucaryotes. Tiré de (Hillis & Dixon, 1991).

Les régions de l'ADN ribosomal évoluent de différentes manières, ce qui permet de générer des données informatives pour presque toutes les questions en classification phylogénétiques. En effet, les deux séquences d'ARNr 18S et 16S sont parmi les séquences qui évoluent lentement chez les espèces vivantes, ce qui les rendent très utiles pour apercevoir les événements évolutifs, en comparaison, les séquences 23S et 28S du gène nucléaire d'ARNr semblent avoir subi beaucoup plus de variations au cours de l'évolution. L'alignement des séquences d'ARNr 16S et 18S fournissent peu de sites variables, ce qui constitue un gène relativement essentiel pour la comparaison des phylums entre les eucaryotes (Hillis & Dixon, 1991).

Outils de recherche et d'alignement de la séquence d'ARNr 18S

De nombreux outils ont été développés pour la prédiction des séquences d'ARNr en raison de leur importance en phylogénie (Van de Peer & De Wachter, 1997). On peut, entre autres, mentionner RNAmmer (Lagesen et al., 2007) et Infernal (Nawrocki, Kolbe, & Eddy, 2009).

Pour chercher les séquences d'ARNr, on procède à une analyse comparative grâce à l'identification des séquences homologues dans le génome. En évolution, les gènes peuvent être regroupés en forme d'un arbre phylogénique grâce à leurs similarités en séquence. Précisément, c'est la fonction du gène qui agit comme une contrainte évolutive. En effet, si le gène est fonctionnellement important, il y a beaucoup de plus de chances de survivre à la prochaine génération et l'inverse est également vrai. Le niveau de conservation des séquences est donc très souvent relié à leur importance fonctionnelle (Eric Paul Nawrocki, 2009).

La technique utilisée pour comparer les séquences est l'alignement. Cela se fait par l'identification des régions similaires entre les séquences homologues. Malheureusement,

l'utilisation de l'alignement pour la comparaison semble moins puissante pour identifier des séquences homologues d'ARN dans des organismes distants (Eric Paul Nawrocki, 2009). En effet, pour identifier une similarité qui est statistiquement significative, l'outil balstn requière une séquence qui doit avoir entre 60% et 65 % d'identité (Eric Paul Nawrocki, 2009). Cette technique est une étape cruciale pour inférer le modèle évolutif de substitutions basé sur la qualité de l'alignement. En raison de la conservation de la séquence ARNr 18S et surtout la structure secondaire qui est la plus conservé chez les familles d'ARN car elle fait partie de leurs fonctions. SSU-ALIGN est la meilleure application à utiliser pour avoir un alignement local ou global d'un haut niveau. En effet, elle utilise des séquences consensus primaires et secondaires pour l'alignement des séquences d'ARN. L'application utilise les profils d'HMM pour la recherche des familles d'ARN afin de renforcer le signal et exploiter la conservation de la structure secondaire d'ARN. Cet outil utilise les modèles de covariance (MC), un modèle probabiliste (voir le schéma de la figure 2.13). Ce qui rend l'application utile et robuste pour aussi la recherche d'homologie. Malgré la puissance de calcul des MC, la recherche sur les bases de données et l'alignement des ARN est extrêmement lente à cause de la complexité algorithmique élevée (Eric Paul Nawrocki, 2009).

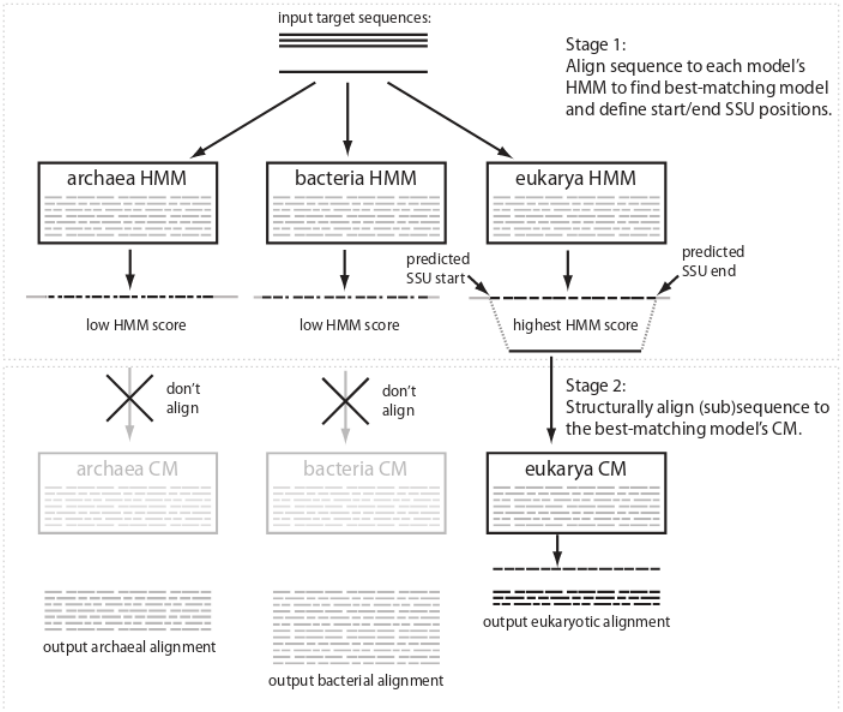

Figure 2.13 : Schéma montrant les étapes d'alignement de SSU-align. Tiré de (Eric Paul Nawrocki, 2009).

3.2.3.4 Modèle évolutif

Le modèle évolutif décrit les processus évolutifs par lesquels les séquences changent au cours du temps, suite à de multiples événements de substitution (Page & Holmes, 2009; Patwardhan et al., 2014). Cependant, ce n'est pas possible de déduire un véritable processus évolutif par observation du nombre de substitutions car un nucléotide identique peut avoir plusieurs mutations (A peut avoir subi A -> T ->A) ou des mutations parallèles (mutation T à A peut avoir réellement subi T -> G -> C -> A). Ce fait réel masque l'estimation des véritables distances évolutives entre les séquences connu sous le nom d'homoplasie. Afin de corriger et de générer un véritable arbre évolutif, des modèles statistiques ont été créés (Patwardhan et al.,

2014). Nous allons décrire les modèles évolutifs, parmi eux, le premier modèle de substitution Jukes-Cantor et le modèle CAT-GTR utilisé lors de notre étude.

Par exemple, le modèle de Jukes-Cantor (JC) créé par Jukes et Cantor en 1969 (Jukes & Cantor, 1969) suppose que les substitutions de purines (A,G) et celles de pyrimidines (C,T) peuvent se produire de manière équiprobable. Ce modèle ne peut par contre être appliqué que sur des séquences d'organismes proches (Patwardhan et al., 2014).

Le modèle généralisé réversible dans le temps (GTR) créé par Tavaré en 1986 (Tavaré, 1986) suppose que les sites évoluent de manière indépendante et que chacun d'eux subissent des substitutions décrites par une chaîne de Markov et réversibles dans le temps. Ce modèle inclut un paramètre qui donne une fréquence inégale pour les quatre nucléotides et un taux distinct pour les six substitutions de nucléotides par paire (Jia, Lo, & Ho, 2014).

Le modèle CAT a été créé par l'équipe d'Hervé Philippe (Nicolas Lartillot & Philippe, 2004). Ce dernier utilise un modèle bayésien qui assume que l'évolution des sites protéiques se fait par des processus (ou classes) distincts. Ces classes diffèrent par les fréquences d'équilibre sur les 20 AA. Les variables libres du modèle sont l'association de chaque site à une classe donnée qui se fait par le processus de Dirichlet [8].

Le modèle CAT-GTR est « un processus de Dirichlet de profils de fréquences d'équilibre combinés à des taux de change généraux » (Nicolas Lartillot, Lepage, & Blanquart, 2009). C'est un très bon modèle pour les données d'ARN et d'ADN. Comparé aux autres modèles,

[8] Famille de processus stochastiques dont les réalisations sont des distributions de probabilité, appliquée dans l'apprentissage automatique.

CAT-GTR semble plus robuste contre les artefacts de longue attraction de branche (Nicolas Lartillot et al., 2009; Nicolat Lartillot, 2018).

3.2.3.5 Méthodes d'inférences

a) - méthodes de distance

Ces méthodes se font par une approche phénétique, c'est-à-dire par un recensement du nombre de différences entre les séquences (estimation de la similarité des séquences) en établissant une matrice de distance (Béatrice Roure, 2011). Pour un petit nombre de taxons, il est possible de calculer l'arbre de distance optimal. Par contre, pour un grand nombre d'organismes, les méthodes heuristiques sont plus utilisées. Mentionnons ici quelques méthodes de distance : (i) Neighbour Joining (NJ) (Saitou & Nei, 1987) qui est une méthode de regroupement qui minimise l'évolution d'un arbre (Page & Holmes, 2009) et (ii) Unweighted Pair Group Method with Arithmetic Mean (UPGMA), une des rares méthodes de regroupement qui construit un arbre ultra métrique (Page & Holmes, 2009). La méthode UPGMA fonctionne par itération : « au début, on suppose que chaque espèce est un regroupement à part entière; on joint les deux groupes les plus proches et on recalcule la distance moyenne des deux ; on répète le processus jusqu'à ce que toutes les espèces soient connectées dans un seul regroupement. » (Yan Li, 2018). Les méthodes de distance sont néanmoins limitées puisqu'elles ne parviennent pas à détecter l'origine de la similarité (Béatrice Roure, 2011). La transformation directe des séquences à leur distance peut engendrer une perte d'information. Par exemple, il est impossible de catégoriser les sites dans un arbre ou de tracer l'évolution des sites individuels (Page & Holmes, 2009).

D'autres méthodes statistiques solides ont par la suite été développées pour faire en sorte qu'elles tiennent compte de plusieurs paramètres évolutifs afin d'obtenir l'arbre le plus adéquat (Nicolas Lartillot, Rodrigue, Stubbs, & Richer, 2013; Stamatakis, 2014).

b) - Méthodes probabilistes

1 - Maximum de vraisemblance

Au début des années 1960, le maximum de vraisemblance a été introduit pour la première fois par Edwards et Cavalli-Sforza en phylogénie (Huelsenbeck & Rannala, 1997). La vraisemblance est la probabilité d'observer les données (séquences génétiques) en fonction de certains paramètres qui forment un modèle évolutif. La fonction $L(\theta|D)$ est la « vraisemblance des paramètres θ sachant les données D, proportionnelle à $P(D|\theta)$ est une probabilité des données D sachant les paramètres θ » (Damien Aubert, 2017). On tient compte du fait que les mutations par substitution sont des événements aléatoires. Dans l'application de la vraisemblance à la phylogénie, « il s'agit de trouver une topologie d'un arbre, les longueurs des branches et les paramètres du modèle mutationnel qui maximisent la probabilité d'observer les séquences disponibles telles qu'elles ont pu être alignées » (Damien Aubert, 2017). Cependant, elle présente une limite puisqu'il peut y avoir des problèmes d'alignements des séquences, ce qui engendre des fausses topologies avec une forte vraisemblance. De plus, cette méthode ne peut pas être liée à un algorithme efficace qui détecte avec certitude les meilleurs arbres sur toutes les topologies possibles car ce sont des problèmes NP-complets (non déterministe polynomial), ce qui signifie qu'il n'y a pas d'algorithme efficace pour trouver la meilleure solution (Page & Holmes, 2009). Ceci fait donc en sorte qu'on ne peut être assuré de trouver un arbre optimal (Damien Aubert, 2017).

2 - Inférence bayésienne

C'est en 1996 que la méthode d'inférence bayésienne a été introduite et appliquée en phylogénie par trois groupes indépendants (S. Li, Pearl, & Doss, 2000; Mau, Newton, & Larget, 1999; Rannala & Yang, 1996) (Frédéric Delsuc. ; Emmanuel Douzery, 2004). Cette méthode est basée sur le calcul des probabilités postérieures d'un arbre qui peut être interprété comme la probabilité à ce qu'un arbre soit « correct » (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001). La probabilité provient du théorème de Bayes :

$$\mathrm{P}(arbre|données) = \frac{\mathrm{P}(\text{ données } | \text{ arbre }) \times \mathrm{P}(\text{arbre})}{\mathrm{P}(\text{données})}$$

P(arbre) est la probabilité à priori ; P(données | arbre) est la vraisemblance (probabilité d'obtenir les données observées si on suppose que l'arbre est vrai) ; P(arbre | données) est la probabilité *a posteriori* (probabilité que l'arbre soit vrai une fois qu'on tient compte des nouvelles données) ; P(données) est la probabilité d'observer les données (Huelsenbeck et al., 2001).

Ces probabilités *a posteriori* ne peuvent pas être calculées analytiquement mais possible d'en établir une approximation par un échantillonnage aléatoire. En effet, la méthode d'échantillonnage de Monte-Carlo par chaines de Markov (MCMC pour Markov Chain monte Carlo) a donc été implémentée afin de maximiser le calcul. L'un des algorithmes les plus utilisés est Metropolis-Hastings (MH) (Damien Aubert, 2017). Le principe de cet algorithme est « de parcourir le graphe E (la densité des probabilités *a posteriori*) de façon aléatoire mais intelligente, en favorisant les arêtes qui font diminuer V, de temps en temps il autorise à augmenter le V pour qu'il ne reste pas bloquer dans un minimum local» (Alouges & Gerin, 2018). Cependant cet algorithme présente des limites dans la convergence des chaines. En

effet, afin d'estimer l'exactitude de la distribution de probabilités *a posteriori*, il faut avoir un nombre suffisant de générations de MCMC (Frédéric Delsuc. ; Emmanuel Douzery, 2004). Cela ne se fait pas en un temps raisonnable car il a plusieurs pics de densité de probabilité isolés dans l'espace des arbres et l'échantillonneur peut rester coincé sur un optimum local. Pour répondre à cette problématique, l'algorithme MH utilise un couplage de MCMC : Metropolis-Coupled Markov chain Monte Carlo, MCMCMC ou MC^3. Cet algorithme, en plus de la chaine principale dite froide, utilise d'autres chaines d'une distribution aplatie des densités de probabilités dites chaudes (Damien Aubert, 2017). Ces chaines chaudes peuvent être distribuées librement sur les pics des chaines froides, ce qui permet à l'échantillonneur d'effectuer une marche sur les autres pics et de trouver l'optimum global. MC^3 est limité par un coût très élevé en termes de calcul, mais est tout de même une bonne méthode de résolution de problèmes dans le domaine de la reconstruction phylogénétique (Damien Aubert, 2017).

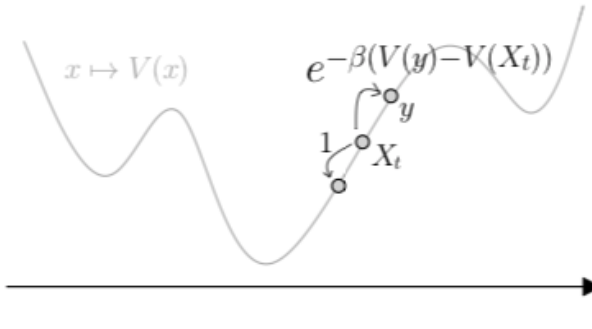

Figure 2.14 : Principe de l'algorithme MH. Tiré de (Alouges & Gerin, 2018).

L'application Phylobayes utilise des alignements nucléotidiques et protéiques pour la construction des arbres phylogénétiques par échantillonnage MCMC. « PhyloBayes se distingue principalement par l'utilisation de méthodes non paramétriques pour la modélisation de l'évolution de la séquence spécifique au site » (Nicolas Lartillot, Blanquart, & Lepage,

2018) . Pour les données protéiques et nucléotidiques, l'outil utilise le modèle GTR (Nicolas Lartillot et al., 2009).

Le programme fait appel à plusieurs sous-programmes (pb, bpcomp), le principal « pb » utilise l'alignement des séquences comme « input » pour produire un arbre, et ce, par la production d'une série de points (ou valeurs) pour tous les paramètres qui sont les longueurs de branches, topologie de l'arbre etc… Ces valeurs sont tirées par la distribution *a posteriori* qui définissent une chaine MCMC. Le nombre de mise à jour de la topologie d'un arbre est enregistré en tant qu'une génération. Un cycle est composé d'un nombre variable de générations qui dépendent du modèle utilisé. Afin d'estimer le nombre de générations effectuées par un échantillonneur MCMC, l' « output » du sous-programme « pb » sont des fichiers dont un a l'extension « .trace » indique l'évolution des générations par le log-semblable en fonction du temps qui désigne quand le profil atteint un état stationnaire. Ce paramètre informe le nombre de générations qui prends pour atteindre l'état stationnaire qui sera utilisé pour éliminer la phase initiale dite « burn-in » par le sous-programme « bpcomp ». En effet, « bpcomp » compare les distributions postérieures de deux chaines parallèles ou plus dans le but d'une bonne estimation de générations. Il générera l'écart le plus important (maxdiff) et le moyen (meandiff) observés sur toutes les bipartitions. La sortie de « bpcomp » est un arbre de consensus créé en regroupant tous les arbres de toutes les chaines ainsi un autre fichier indiquera la qualité de l'arbre calculé par le « maxdiff » (voir tableau) (Nicolas Lartillot et al., 2018).

Maxdiff	Interprétation
Inférieur 0.1	Très bon.
Inférieur à 0.3	Acceptable.
Entre 0,3 et 1	Échantillon pas assez grand.
Égale à 0	Pas bon, au moins une des chaines est bloquée dans un maximum local.

Tableau 2.3 Interprétation de « maxdiff ». Tiré de (Nicolas Lartillot et al., 2018)

Chapitre 3

Article

Pervasive convergent evolution and extreme phenotypes define chaperone requirements of protein homeostasis

Yasmine Draceni [1], Sebastian Pechmann[1]

[1] Department of Biochemistry, Université de Montréal, Montréal, QC, H3T 1J4, Canada

Target journal of submission: Proceedings of the National Academy of Sciences (PNAS).

Keywords: evolution, chaperones, phylogenetic, proteostasis

Author contributions

YD developed, implemented, and carried out all computational analyses and prepared the initial manuscript and figures. **SP** devised and supervised the project and assisted in the preparation of the final manuscript and figures. **All authors** contributed to the interpretation of results and approved the final version.

Acknowledgements

We are grateful to Emmanuel Noutahi and Simon Laurin-Lemay for assistance with the PhyloBayes calculations, and the members of the Pechmann Lab for helpful discussions. This work was funded through financial support from an NSERC Discovery Grant and the Université de Montréal. Computations were in part performed on supercomputers managed by Calcul Québec and Compute Canada that are funded by the Canada Foundation for Innovation (CFI), Ministère de l'Économie, des Sciences et de l'Innovation du Québec (MESI) and le

Fonds de recherche du Québec – Nature et technologies (FRQ-NT). SP holds the Canada Research Chair in Computational Systems Biology.

3.1 Abstract

Maintaining protein homeostasis is an essential requirement for cell and organismal viability. An elaborate regulatory system within cells, the protein homeostasis network, safeguards that proteins are correctly folded and functional. At the heart of this regulatory system lies a class of specialized protein quality control enzymes called chaperones that are tasked with assisting proteins in their folding, avoiding aggregation, and degradation. Failure and decline of protein homeostasis is directly associated with conditions of aging and aging-related neurodegenerative diseases such as Alzheimer's and Parkinson's. However, it is not clear what tips the balance of protein homeostasis and leads to onset of aging and diseases. Here, we present a comparative genomics analysis of protein homeostasis in eukaryotes and report general principles of maintaining protein homeostasis across the eukaryotic tree of life. We find, as previously reported, a strong correlation between the size of eukaryotic chaperone networks and size of the genome, thus suggesting pervasive convergent evolution. Notably, this observed correlation is distinct for different species kingdoms. Importantly, organisms with pronounced phenotypes such as *Northobranchius furzeri*, the African killifish and shortest-lived vertebrate, or *Heterocephalus glaber*, the naked mole rate and longest-lived rodent, clearly buck this trend. *N. furzeri* as widely used model for fragile protein homeostasis is found to be chaperone limited, while *H.glaber* as extra robust organism is characterized by above average number of chaperones. Our work thus indicates that the balance in protein homeostasis may be a key variable in explaining organismal robustness. Finally, our work

provides an elegant example of harnessing the power of evolution and comparative genomics to address fundamental open questions in biology with direct relevance to human diseases.

3.2 Introduction

Keeping one's proteome properly folded and functional through varying conditions and stresses is of critical importance for cellular and organismal survival (Chiti & Dobson, 2006). Conversely, the decline of the cell's capacity to keep its proteins in their correct shape, i.e. to maintain protein homeostasis, is a central hallmark of aging and onset of aging-associated diseases (Kaushik & Cuervo, 2015). Naturally, there is intense interest to better understand the principles of successful protein homeostasis as well as the origins of its failures. Cells maintain protein homeostasis through a complex regulatory network that integrates protein synthesis, folding, degradation, and trafficking pathways (Powers, Morimoto, Dillin, Kelly, & Balch, 2009). The central players of the protein homeostasis network are a class of specialized protein quality control enzymes called chaperones (Brehme et al., 2014; Y. E. Kim, Hipp, Bracher, Hayer-Hartl, & Hartl, 2013). Chaperones assist proteins in their folding, protect them from aberrant aggregation while promoting functional assembly, and, if needed, sequester and target them for degradation (Y. E. Kim et al., 2013, p. 20).

Eukaryotic genomes often contain more than ~50-200 different chaperone genes that classify into distinct families based on the structure and function of the encoded proteins. As the best-characterized heat shock protein, Hsp90 type chaperones are the main stress responders that stabilize and refold partially unfolded and stress-denatured proteins (McClellan et al., 2007;

Taipale, Jarosz, & Lindquist, 2010). Hsp90 chaperones interact with large and diverse sets of client proteins (McClellan et al., 2007), are highly expressed, and usually encoded by only few genes (B. Chen, Zhong, & Monteiro, 2006) that are highly activated under unfavorable conditions (Albanese, Yam, Baughman, Parnot, & Frydman, 2006). The family of Hsp70 chaperones are key enzymes in determining the fate of newly made proteins (Sebastian Pechmann, Willmund, & Frydman, 2013): Hsp70 often assist in the *de novo* folding of nascent polypeptides (Willmund et al., 2013), their translocation, and disaggregation (Nillegoda & Bukau, 2015). Many Hsp70s are thus transcriptionally coupled to protein biosynthesis (Albanese et al., 2006). Hsp40 chaperones primarily act as co-chaperones and nucleotide exchange factors to Hsp70s, thus providing specificity in guiding protein folding, assembly, disassembly, and translocation (Nillegoda et al., 2017; Qiu, Shao, Miao, & Wang, 2006; Walsh, Bursać, Law, Cyr, & Lithgow, 2004). Small heat shock proteins (sHsp or Hsp20) stabilize partially denatured proteins and prevent their aggregation (Sun & MacRae, 2005), while enzymes from the Hsp100 chaperone family are primarily involved in disaggregation and proteolysis (Leitner et al., 2012; Yam et al., 2008). Hsp60s encode protein subunits of eukaryotic chaperonins, which are heteromeric protein complexes that provide a fully enclosed protective cavity for the folding of select proteins (Leitner et al., 2012; Yam et al., 2008). Together with their individual functional specializations, the ensemble of all chaperones in the cell provides a powerful network to control the quality of proteins.

Chaperones generally bind to insoluble, sticky proteins that are at risk of aggregation (Gian Gaetano Tartaglia, Dobson, Hartl, & Vendruscolo, 2010). However, unlike in bacteria where each chaperone family is represented by only very few genes, chaperone genes in higher

eukaryotes strongly increase in numbers. This is particularly true for chaperones of the Hsp70 and Hsp40 families that can be found with strongly increasing diversity in eukaryotes (Qiu, Shao, Miao, & Wang, 2006). The reasons why so many different types of chaperones are necessary to keep the proteome folded is far from understood. In specifically interacting with select pools of client proteins, chaperones are likely key regulatory elements of cellular networks (Sőti, Pál, Papp, & Csermely, 2005). Moreover, because both biosynthesis and maintenance as well as the ATP-dependent activity of chaperones comprise substantial energetic costs, the increasing diversity in eukaryotic chaperone families has likely arisen under selection and directly reflects the requirements of managing increasingly complex proteomes (Santra, Farrell, & Dill, 2017).

To this end, the analysis of dynamically regulated proteomes in response to stress (Sala et al., 2017) as well as the comparative analysis of related organisms (Powers & Balch, 2013) offer the opportunity to gain fundamental insights into how the balance between the size and composition of the chaperone network as functional core of cellular protein homeostasis and the composition of the proteome is maintained (Roth, 2011). For instance, a general correlation between the size of the chaperone network and number of protein coding genes could be observed for 16 eukaryotic genomes (Powers & Balch, 2013). Malignant tumor growth in cancers is aberrantly balanced by dynamically adjusted chaperone expression profiles (Hadizadeh Esfahani, Sverchkova, Saez-Rodriguez, Schuppert, & Brehme, 2018). And even viral infections impose specific requirements on the host cell chaperone network to sustain their replication (R. Geller, 2007; Ron Geller, Pechmann, Acevedo, Andino, & Frydman, 2018; Taguwa et al., 2015).

Notably, chaperones are some of the most ancient and evolutionarily conserved components of the cell. Accordingly, evolutionary approaches offer a great opportunity to learn more about the functioning and organization of chaperone networks and cellular protein homeostasis (D. Bogumil, Alvarez-Ponce, Landan, McInerney, & Dagan, 2014; David Bogumil, Landan, Ilhan, & Dagan, 2012; Kultz, 2003; Press et al., 2013). For instance, the evolutionary history of the Hsp90 family suggests multiple independent duplication and gene loss events (Chen et al., 2006), reflecting a perpetual challenge to stay in balance with an equally evolving proteome. In turn, the expansion of proteomes (Caetano-Anolles, Kim, Mittenthal, & Caetano-Anolles, 2011) and encoded protein networks (Korcsmaros, Kovacs, Szalay, & Csermely, 2007; S. Pechmann & Frydman, 2014) themselves is promoted by chaperones. Within proteomes, a dominant evolutionary force is given by selection to avoid protein aggregation (Dobson & Dobson, 1999). This manifests in diverse strategies to mitigate the risk of aggregation through aggregation resistant sequences (Monsellier, Ramazzotti, Taddei, & Chiti, 2008; G. G. Tartaglia, Pellarin, Cavalli, & Caflisch, 2005) and evolutionary design principles that protect proteins from aggregation (Beerten et al., 2012; S. Pechmann et al., 2009), strategies to temporarily sequester aggregation prone proteins (Wallace et al., 2015), or increased protein quality control for aggregation prone proteins (Bershtein, Mu, Serohijos, Zhou, & Shakhnovich, 2013; Sebastian Pechmann & Vendruscolo, 2010; Willmund et al., 2013). The combinatorics of these diverse strategies hints at the complexity of understanding protein homeostasis.

Evolution offers striking examples of innovation to counter varying challenges, conditions, or constraints. In particular, the study of extreme phenotypes offers fascinating insights into the trade-offs underlying organismal fitness. At the protein level, thermophilic organisms counter the challenge of thermodynamic protein stability at elevated temperatures with reduced surface hydrophobicity to avoid aggregation yet an increase in buried sequence hydrophobicity to stabilize the protein structural core (Greaves & Warwicker, 2007; Thangakani, Kumar, Velmurugan, & Gromiha, 2012). Similarly, specifically evolved ice-binding proteins allow species to thrive at sub-zero temperatures (Davies, 2014). At the cellular level, the *Dictyostelium discoideum* proteome with an extremely high number of aggregation-prone prion domains is balanced by an increase in Hsp100 disaggregase capacity (Malinovska, Palm, Gibson, Verbavatz, & Alberti, 2015). Recently established powerful model systems for fragile or especially robust protein homeostasis include *Northobranchius furzeri* (Reichwald et al., 2015; Valenzano et al., 2015), the African Killifish, and *Heterocephalus glaber* (Keane et al., 2014), the naked mole rat. *N.furzeri* is the shortest-lived vertebrate. Within a life expectancy of only 4 – 6 months (Valenzano et al., 2015), *N.furzeri* progresses through many of the hallmarks associated with human aging including pronounced loss of protein homeostasis. The naked mole rat *H.glaber* in turn is the longest living rodent with a life expectancy of over 30 years (Keane et al., 2014). This stands in stark contrast to strongly related rodents such as *Mus musculus* that only live for around 3 years (Keane et al., 2014). In both cases the genetic origins of the comparably short and long lifespans respectively are far from understood. However, in all cases, aging is characterized by the loss of protein homeostasis (Schneider & Bertolotti, 2015) and the accumulation of misfolded and

aberrantly aggregated proteins (David, 2012), underlining the importance of better understanding how cells keep their proteomes in balance.

Here, we present a comparative genomics analysis of protein homeostasis across the eukaryotic tree of life. By significantly expanding an initial analysis of 16 eukaryotic genomes (Powers & Balch, 2013) to the analysis of over 200 eukaryotes, we report general principles of keeping the balance of successful protein homeostasis. A strong correlation between the size of eukaryotic chaperone networks and size of the genomes suggests pervasive convergent evolution that follows slightly distinct trends for different kingdoms. *N. furzeri* as widely used model for fragile protein homeostasis is found to be chaperone limited, while *H.glaber* as especially robust organism is characterized by above average numbers of chaperones. Our work thus indicates that the balance in protein homeostasis may be a key variable in explaining organismal robustness.

3.3 Results

To investigate principles of the evolution of protein homeostasis in eukaryotes, we first sought to collect a comprehensive high-quality dataset of annotated proteomes that we could analyze in the context of their phylogenetic relationships. Phylogenetic models across a highly diverse set of species are best computed based on a single marker (Patwardhan et al., 2014), in eukaryotes usually the highly conserved 18S ribosomal RNA (rRNA) gene. Of note, the assembly and annotation state of available genomic data differs widely. Despite its central importance for phylogenetic analyses, the rRNA locus is not always annotated in early draft genomes, likely due to difficulties with mapping the many repetitive sequence elements of the intra-genic linker regions (Noller, 1984). Therefore, we tested different approaches to efficiently identify 18S rRNA sequences in sequenced genomes: the popular heuristic sequence similarity search algorithm BLAST (Camacho et al., 2009) that is fast but not very accurate, and the hidden Markov model (HMM) based algorithm RNAmmer (Lagesen et al., 2007) that is very accurate but slow. Finally, to circumvent these trade-offs, we developed a pipeline based on available algorithms to readily identify 18S rRNA sequences in genomes (Fig. 3.1A). Specifically, we created artificial 50nt-long short reads from the annotated and validated rRNA sequences of the model organisms *C.elegans*, *D.melanogaster*, *H.sapiens*, *M.musculus*, and *S.cerevisiae*. that were aligned with HISTA2 (D. Kim, Langmead, &

Salzberg, 2015) to the target genome to identify the candidate rRNA locus that contains the 18S rRNA sequence (*see Methods*). RNAmmer (Lagesen et al., 2007) was subsequently used to identify the exact 18S rRNA sequence from the candidate rRNA locus.

We were able to identify 18S rRNA sequences by BLAST in only 19%, and by RNAmmer in only 25% of the eukaryotic genomes from the RefSeq databse (Fig. 3.1B; Table 3.S1). Moreover, BLAST yielded many false positive hits that upon closer inspection did not align well with validate 18S rRNA sequences. We concluded that BLAST was not an appropriate tool to discover rRNA sequences. The software RNAmmer is highly accurate in the identification of rRNA but too slow to efficiently scan full genomes. Therefore, we could only obtain 18S rRNA with RNAmmer from genomes where the rRNA locus had already been annotated. With our strategy to use the short-read aligner HISAT2 and synthetic rRNA reads from a mix of source organisms to map the best candidate rRNA locus followed by the accurate identification of the 18S rRNA with RNAmmer we could expand our dataset to 62% of the initially downloaded genomes (Fig. 3.1B; Table 3.S1). The missing rRNA sequences in the remaining genomes likely suffer from challenge to resolve the repetitive sequences in the rRNA cassette in early genome assembly states. Our results suggest that the combination short-read alignment and RNAmmer can be a very efficient manner to identify rRNA sequences in genomes. Additional conservative data quality control (*see Methods*; Fig. 3.1C; Fig. 3.S1A, B) resulted in a set of 216 eukaryotes spanning Fungi, Protista, Metazoa, and Plantae, i.e. all four eukaryotic kingdoms of life, for which we computed a phylogenetic tree (Fig. 3.1D). The obtained phylogenetic model was supported by a good *maxdiff* = 0.27 between three independent trees and found in very good agreement with taxonomy annotations

(Fig. 3.1D). Moreover, the consistent clustering of species within their kingdoms quantified to clear differences in short pairwise evolutionary distances between species within their cluster compared to longer distances to the outside of their cluster (Fig. 3.S1C). Protists were found the least separated from the other kingdoms, which may also be the result of low representation in our dataset. In contrast, both animals and plants partitioned into well-defined and distinct clusters (Fig. 3.1D).

We next wanted to better understand principles of the evolution of chaperone networks as core components of the protein homeostasis network across this eukaryotic phylogeny. All occurrences of nuclear genome encoded chaperones of the Hsp20, Hsp40, Hsp60, Hsp70, Hsp90, and Hsp100 families were identified (*see Methods*) and mapped onto the phylogenetic tree (Fig. 3.2). Strikingly, patterns of the counts of chaperone genes directly correlate with similarities and divisions in the phylogeny. Increasing organism complexity is accompanied by increasing diversity in all chaperone families (Fig. 3.2). Especially visible is the clear transition from animals to the polyploid plant genomes (Fig. 3.2). The shift to larger chaperone networks for more complex genomes is noticeable for all chaperone families, but especially standing out are increased counts for Hsp20, Hsp40, and Hsp70 type chaperones (Fig. 3.S2). Within the different kingdoms, plants show by far the largest variance in the composition of their chaperone networks (Fig. 3.2, 3.S2). While the composition of eukaryotic chaperone networks appears in general to be strongly correlated, individual species are clearly standing out by above or below average numbers in individual chaperone families (Fig. 3.2). Taken together, our analysis of the chaperone networks across eukaryotes suggests fundamentally conserved similarities with pronounced exceptions.

We next systematically evaluated these observed (dis)similarities in the composition of eukaryotic chaperone networks as function of their evolutionary distance. Pairwise chaperone network similarity was quantified by the correlation coefficient between the chaperone counts in the six protein families and compared to the evolutionary distances obtained from the branch lengths of the phylogenetic tree. To highlight organisms that differ the most from the rest, we computed for each species the average of its pairwise correlation coefficients to the chaperone profiles of all other species, and similarly the average of the evolutionary distances to all other species (*see Methods*). Focusing on the most populated clusters, i.e. the fungi (Fungi), animals (Metazoa), and plants (Plantae), we observed a general strong correlation of the composition of the chaperone networks independent of average evolutionary distance (Fig. 3.3A). In the Fungi and Plantae groups almost all species had average correlation coefficients above 0.9, thus suggesting that the chaperone networks in eukaryotes are generally composed of similar relative numbers of Hsp20, Hsp40, Hsp60, Hsp70, Hsp90, and Hsp100 type chaperones. Because these different chaperones perform overlapping yet specialized functions it is not surprising that all of them are proportionally present. Moreover, the observation that the generally very high similarity in chaperone profiles across eukaryotes is found to be independent of evolutionary distance (Fig. 3.3A) indicates a fundamental selective pressure on proteome composition, likely through convergent evolution.

The fungus with the lowest average correlation coefficient was *Rizophagus irregularis*, a symbiotic fungus used in soil agriculture that belongs to a different taxonomy division than the rest of the fungi in our data. The plants with the lowest average correlation coefficient in our

dataset were *Rosa chinesis* and *Capsicum annuum*, two highly domesticated plants (Fig. 3.3A). In contrast, the cluster of animal species in our dataset exhibited clearly more diversity. Specifically, the Pacific oyster *Crassostrea gigas* was found to have the most dissimilar chaperone profile with an average correlation coefficient of only 0.39, followed by the sea urchin *Strongylocentrotus purpuratus* (average correlation = 0.64). We decided to analyze more in-depth the chaperone networks of the two most dissimilar animals found by this analysis, *C.gigas* and *S.purpuratus*, together with two organisms of known phenotype of interest, *N.furzeri* and *H.glaber* (Fig. 3.3B, C). Upon closer inspection it became apparent that *C.gigas* had an extraordinary high count of Hsp70 chaperones at over 80 (Zhang et al., 2012) while the average across the studied animal genomes was only around 15 (Fig. 3.3B,C). However, *C.gigas* only has 6 Hsp100 (Fig. 3.S3). The sea urchin *S.purpuratus* was equally characterized by a high number of over 40 Hsp70 chaperones (Fig. 3.3B,C) but also one of the highest observed count of Hsp100 (Fig. 3.S3). Strikingly, the naked mole rat *H.glaber* displayed one of the highest counts of Hsp40 chaperones and ranked amongst the highest counts of both Hsp70 and Hsp90 chaperones of all animals studied (Fig. 3.3B,bC). In contrast, the African killifish *N.furzeri* stood out for one of the lowest observed counts of Hsp90 chaperones among the group of animals (Fig. 3.3B, C). Taken together, all examples that represent extreme phenotypes were indeed also characterized by extreme counts of specific chaperone types as compared within the group of animal genomes analyzed.

Because chaperones preferentially bind to aggregation prone proteins, we next quantified the aggregation propensity of the proteomes across the phylogeny (Fig. 3.4A). The distributions of per-protein predicted aggregation propensity were represented by the characteristic median

and lower and upper percentiles and clearly correlated well for many closely related species (Fig. 3.4A). On average, fungi and animals showed a higher aggregation scores than plants (Fig. 3.4B). The combination of both higher chaperone counts and lower aggregation scores in plants hints at the complexities of their life styles. Remarkably, the median aggregation score of the proteomes of *C.gigas*, *S.purpuratus*, *H.glaber*, and *N.fuzeri* fell right on the center of the corresponding distribution of all animals studied, or even placed on the lower end of the 25% percentile (Fig. 3.4C). This suggests that these organisms do not possess proteomes with elevated aggregation scores.

Biological organisms are most beautifully complex systems, and complex phenotypes can rarely be explained by individual observables. To understand how species may balance their proteome with costly chaperone capacity, we next compared the size of the chaperone networks in individual species with the overall aggregation score of their proteomes (Fig. 3.4D). Specifically, we first considered Hsp40, Hsp70, and Hsp90 as the core chaperone network as these are the main chaperones for initial folding, stress response, and disaggregation. A proteome aggregation score was computed as the sum of the individual predicted protein aggregation propensities, thus reflecting the size of the proteome weighted by its propensity to aggregate. Remarkably, we observed a general and very strong linear correlation between the size of the core chaperone network and the proteome aggregation score (Fig. 3.4D) that is much more pronounced that what had previously been reported for a subset of 16 eukaryotes (Powers & Balch, 2013). Strikingly, several organisms clearly diverted from this trend including the ones previously identified. The naked mole rat *H.glaber* was found far to the right of this general trend, suggesting an excess of chaperones compared

to the proteome aggregation propensity and relative to the other animals (Fig. 3.4D). This observation matches very well the remarkable robustness and longevity of *H.glaber*. The Pacific oyster *C.gigas* could also be found far to the right of the general trend (Fig. 3.4D). In contrast, the African killifish *N.fuzeri* placed on the left side of the general correlation (Fig. 3.4D), reflective of its known fragile protein homeostasis. Similarly, *Mus musculus* as one of the shortest-lived mammals equally placed clearly to the left side of this correlation, as did the nematodes *C.elegans* and *C.remanei* (Fig. 3.4D). he sea urchin *S.purpuratus* however fell onto the general trend, suggesting that its expanded chaperone network merely compensates a more aggregation prone proteome and other putative requirements (Fig. 3.4D). Importantly, the same observations could be made when considering all chaperone families (Fig. 3.S4A). Moreover, the analysis of our complete dataset across four kingdoms revealed that in individual species kingdoms follow slightly different trends with a decreasing slope from fungi and protists to animals and finally plants (Fig. 3.S4B). Taken together, our results reveal fundamental principles of keeping protein homeostasis in balance that explain known cases of particularly weak or robust protein homeostasis.

Next to these global trends in the balance between chaperones and proteome, our observations match many fascinating facets of species-specific constraints and adaptations. The sea urchin *S.purpuratus* is an important model organism for developmental biology and characterized by a remarkable lifespan of over a century as well as high fecundity (Cameron, Samanta, Yuan, He, & Davidson, 2009; Sea Urchin Genome Sequencing Consortium et al., 2006). With a high degree of polymorphisms in its population (Cameron et al., 2009) and strong defense systems (Goldstone et al., 2006), *S.purpuratus* contained one of the largest counts of Hsp100

chaperones of all animals analyzed. The proteolytic activity of Hsp100s may assist the complex and sophisticated innate immune system (Smith, 2012) in defending against a plethora of stressors, as well as the biosynthesis and assembly of its pronounced spikes. Another maritime organisms, the Pacific oyster *C.gigas* has evolved unusually high numbers of chaperones to adapt to harsh living conditions and the specific biosynthesis requirements of making the hard oyster shells (Zhang et al., 2012). Also living in the highly stressful intertidal zone, the genome is characterized by a remarkable expansion of Hsp70s (Zhang et al., 2012). Moreover, the oyster shell is formed by a diverse array of matrix proteins through complex assembly and modification processes (Zhang et al., 2012). Hsp70 chaperones frequently function as assembly factors. Thus, next to an extraordinary capability to adapt to stress conditions (Lim et al., 2016; Wang et al., 2019), the large number of Hsp70s may play a crucial role in shell formation. With respect to longevity, the most striking observation from our analyses is the opposing numbers of Hsp90 proteins in *N.furzeri* and *H.glaber*. The short-lived *N.furzeri* has the lowest and the long-lived *H.glaber* the highest number of Hsp90 genes in the group of Animalia. While both proteomes are likely well balanced under normal conditions, the ability to adapt to stress conditions is critical for long-term survival. While *N.furzeri* has strong capacity in the Hsp20 and Hsp40 systems, Hsp70 is equally only present in relatively few copies. The naked mole rate *H.glaber* is often compared to *M.musculus* for their stark contrast in longevity (Rodriguez et al., 2016) and here equally stands out the much higher numbers of Hsp70 and Hsp90 genes in the longer-lived *H.glaber* (Fig. 3.4D). Another short-lived model organisms for aging research, the nematode *C.elegans*, is equally found to be chaperone limited (Fig. 3.4D). The individual living conditions and phenotypes of different organisms are as complex as the protein homeostasis network, that will ultimately have to be

understood at a much more detailed level. However, our comparative genomics analyses reveal fundamental trends of how organisms keep their proteome in balance, and how diversion from this trend correlates with known extreme phenotypes.

3.4 Discussion

We have performed a comparative genomics analysis of protein homeostasis in 216 eukaryotes. Our results suggest strong convergent evolution to maintain the overall composition of eukaryotic chaperone networks across the eukaryotic kingdoms of life. Strikingly, those organisms that divert from the general chaperone profile are often accompanied by specific phenotypic constraints. Some of the shortest-lived species have the lowest numbers of Hsp90 genes, the main stress response chaperone, while especially robust organisms are characterized by high Hsp90 chaperone counts. Our finding that the composition of chaperone networks with proportional numbers of Hsp20, Hsp40, Hsp60, Hsp70, Hsp90, and Hsp100 is generally very stable and independent of evolutionary distance underlines that this evolutionarily ancient machinery strongly co-evolved with the corresponding proteomes. In turn, the finding that the cases with clearly different chaperone networks respond to special circumstances offers a fascinating window into the comparative genomics of protein homeostasis.

A general short-coming of our analysis lies in the inference from the number of nuclear encoded chaperone genes to their protein folding capacity in the cell, which equally strongly

depends on their expression profiles. Nonetheless, from an engineering and systems perspective there are fundamental differences between having few loci that are highly expressed and many gene copies, even if both result in similar protein levels. As organisms age and accumulate mutations, the biosynthesis of chaperones themselves may be negatively affected. By distributing the genetic information across several loci, the system becomes more redundant thus robust (Stelling, Sauer, Szallasi, Doyle, & Doyle, 2004). Equally importantly, diversity in chaperones likely yields a gain in specificity and fidelity in regulation, thus affording more control over adaptive responses to stress. Our results and the results by others suggest that this is a fundamental constraint on managing more complex proteomes of higher eukaryotes, but likely similarly affects organismal fitness and longevity.

As much as our work highlights the power of comparative genomics to shed light onto fundamental open questions in biology, as much does our work reveal current challenges and opportunities for further improvement. Large-scale comparative genomics analyses are strongly dependent on the quality and progress of genome assemblies and annotations, on their own formidable challenges. While final numbers of genes including chaperone encoding genes likely continue to change minimally with progressing assembly and annotation, our numbers are in overall very good agreement with previously published analyses of chaperone systems.

Finally, protein homeostasis is achieved through a truly complex regulatory network and protein folding under chaperone assistance only comprises one central aspect of it (Klaips, Jayaraj, & Hartl, 2018). Many parts of this regulatory system are tightly controlled, and even RNA molecules are chaperoned (Rudan, Schneider, Warnecke, & Krisko, n.d.). An alternative

means to remove misfolded and aggregated proteins is provided by powerful degradation systems whose malfunctioning is equally implicated in organismal aging and disease (Dikic, 2017). The continued explosion of sequenced genomes without doubt will enable to significantly expand evolutionary and comparative genomics efforts to learn about protein homeostasis.

3.5 Materials and Methods

***Code availability*:** Computer code to reproduce all results and analyses is available at github.com/pechmannlab/chapevo

Data sources: *W*e downloaded all eukaryotic genomes from the RefSeq database (O'Leary et al., 2015) that also had corresponding validated proteomes available in the Uniprot database (Apweiler et al., 2004). This yielded a set of 472 eukaryotic genomes and corresponding proteomes. From RefSeq, both annotated RNA sequences and whole genome FASTA files were downloaded. Draft genome assemblies of *Heterocephalus glaber* (Keane et al., 2014) ("naked mole rat") and *Nothobranchius furzeri* (Reichwald et al., 2015; Valenzano et al., 2015) ("African killifish") were downloaded individually. The corresponding protein sequences for each organism were retrieved from the Uniprot database using the provided FASTA files without isoforms, and in the case of *N.furzeri* from the species-specific database (http://nfingb.leibniz-fli.de/) and filtered for one isoform per gene based on gene IDs.

***Identification of chaperones*:** We used the expert-curated profile hidden Markov models (HMMs) from the Heat Shock Protein Information Resource database (Sinha et al., 2012) and the *hmmsearch* module of the Hmmer (S. R. Eddy, 2011) software to identify Hsp20, Hsp40, Hsp60, Hsp70, Hsp90 and Hsp100 chaperones in sequenced and annotated proteomes. All hits of *hmmsearch* with e-values below 10^{-5} were considered.

***Identification of 18S rRNA*:** Three approaches were tested to efficiently find the 18S rRNA gene in genomes. First, annotated 18S rRNA sequences from the well-characterized model organisms *H. sapiens*, *A.thaliana*, *D.melanogaster* and *R.norvegicus* were searched with BLAST (Camacho et al., 2009) against the target genome. Herein, best hits with *blastn* and e-values < 10^{-6} as widely accepted threshold for homology searches (Pearson, 2013) were considered as putative rRNA loci. BLAST is, in generally, insufficient to accurately detect 18S rRNA genes as we obtained many false positive hits. Next, the software RNAmmer (Lagesen et al., 2007) was used to identify 18S rRNA sequences from genomes. RNAmmer is a highly accurate HMM-based software but too slow to search whole genomes. Therefore, 18S rRNA sequences could only be identified from genomes with already annotated rRNA loci. Last, we converted the validated 18S rRNA sequences from *C.elegans*, *D.melanogaster, H.sapiens*, *M.musculus*, and *S.cereviceae* with a sliding window of length 50 into artificial short RNA reads. The choice of source organisms was arbitrary but found to yield robust results. For each target genome an alignment index was built and the short-read aligner HISAT2 (D. Kim et al., 2015) was used to align the mix of rRNA reads form the different species to the full target genome sequence. From the resulting SAM file, the chromosome and locus with the highest coverage was identified, and the start and end positions both extended as candidate rRNA locus. We obtained robust results for extending the candidate locus by 100, 500, or 1500nt. RNAmmer was next used to efficiently identify the coordinates and sequence of the 18S rRNA within this candidate locus.

***Phylogeny of eukaryotes*:** We removed 28 of 18S rRNA sequences that contain "N" characters as they impede accurate evolutionary inference. To construct a high-quality

sequence alignment, the structure-based SSU-align (Nawrocki & Eddy, 2010) program was used as alignment tool. Next, a phylogenetic tree was constructed with PhyloBayes (Nicolas Lartillot et al., 2009, 2013) using the CAT/GTR model and three independent Markov Chain Monte Carlo (MCMC) chains of length 10,000. The first 1000 trees were cast-off as burn-in. Our initial tree showed a largest discrepancy observed across all bipartitions of *maxdiff* = 0.534, thus suggesting under-sampling. Specifically, upon manual inspection of the 18S rRNA alignment it was clear that only few sequences caused the multiple-sequence alignment to expand by an above average number of gaps. To identify sequences that caused the largest number of gap openings, we computed pairwise sequence alignments of all sequences in our 18S rRNA dataset with SSU-align and counted the number of gaps. We considered the '% gaps' as the number of gaps relative to alignment length in pairwise alignments. The 'average % gaps' thus represents the average percentage of gaps a sequence introduces to all other sequences in pairwise alignments. In trying to obtain the best tree, we sought to minimize the *maxdiff* while keeping as many sequences as possible. Thus, phylogenetic trees were computed upon removal of sequences exceeding six different thresholds for the 'average % gaps': 15%, 20%, 25%, 30%, 50%, 100%. The threshold of 15% was found to yield the best results, and sequences with a gap score of > 15% were discarded (Table 3.S2). This yielded a dataset of 216 eukaryotic genomes. The largest discrepancy between three independent chains reduced to *maxdiff* = 0.27, thus indicating a good phylogenetic model (Nicolas Lartillot et al., 2013).

***Taxonomy classification*:** Species taxonomy information was retrieved from the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy) and assigned with the R

package myTAI (Drost, Gabel, Liu, Quint, & Grosse, 2017). We assigned species with missing kingdom annotations to the cluster of "Protists".

***Aggregation scores*:** The propensity of proteins to aggregate was predicted from the protein amino acid sequences with the Tango (Fernandez-Escamilla et al., 2004) software. Tango predicts per-residue aggregation scores that can be summed up to per-protein scores. We considered the distribution of all predicted per-protein aggregation scores as represented through the median and characteristic 25% (Q1) and 75% (Q3) percentiles. Finally, *proteome aggregation scores* were computed as the sum of the individual protein aggregation scores, thus reflecting the size of the proteome weighted by its aggregation propensity.

***Data analysis*:** All data analyses were performed with custom computer code written in Python and R (github.com/pechmannlab/chapevo). Figures were prepared in R including use of the packages *ggplot2* (Wickham, 2016), *ggtree* (Yu, Lam, Zhu, & Guan, 2018; Yu, Smith, Zhu, Guan, & Lam, 2017), and *cowplot* (https://github.com/wilkelab/cowplot).

Figures:

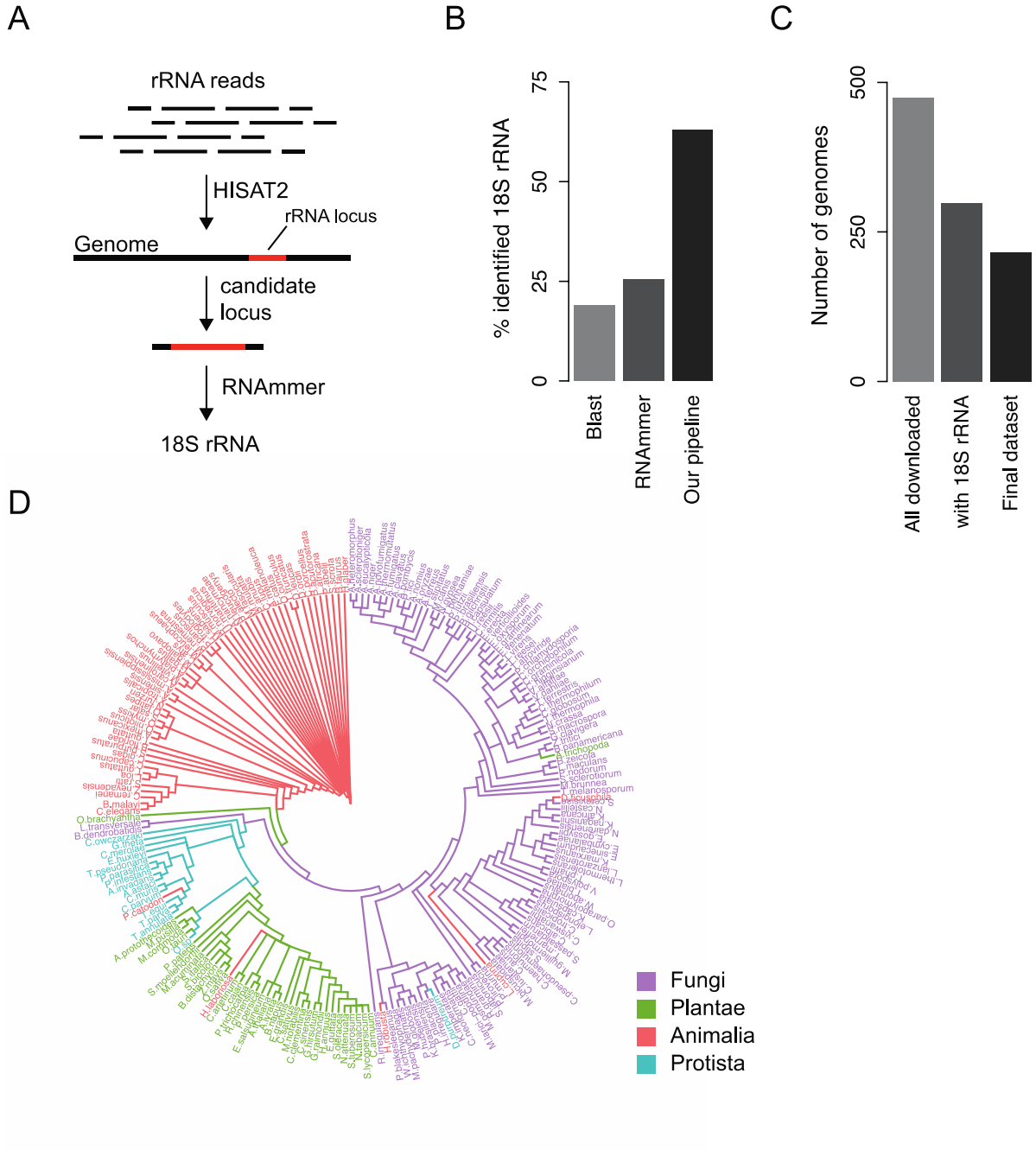

Figure 3.1. Phylogenetic classification of eukaryotes. **A** Pipeline for the efficient identification of 18S rRNA sequences in genomes based on aligning short rRNA reads with HISTA2 and RNAmmer. Mapping of synthetic rRNA reads is used to find the best candidate locus, followed by exact 18S rRNA identification with RNAmmer. **B** Benchmarking the identification of rRNA. The developed pipeline improves the efficient identification of 18S

rRNA sequences in genomes (Table 3.S1). **C** Data quality control of all genomes with genome assemblies in the RefSeq database and validated proteomes in the Uniprot database (all downloaded) and identified 18S rRNA sequences. Only few additional organisms were excluded from further analyses; 28 because their 18S rRNA sequence contained 'N' characters, and the rest because the 18S sequences introduced above average numbers of gaps into the multiple sequence alignment (*see Methods*, Table 3.S2). **D** Eukaryotic tree of life colored by species kingdoms. Strong clustering of species within their kingdoms supports a good phylogenetic model.

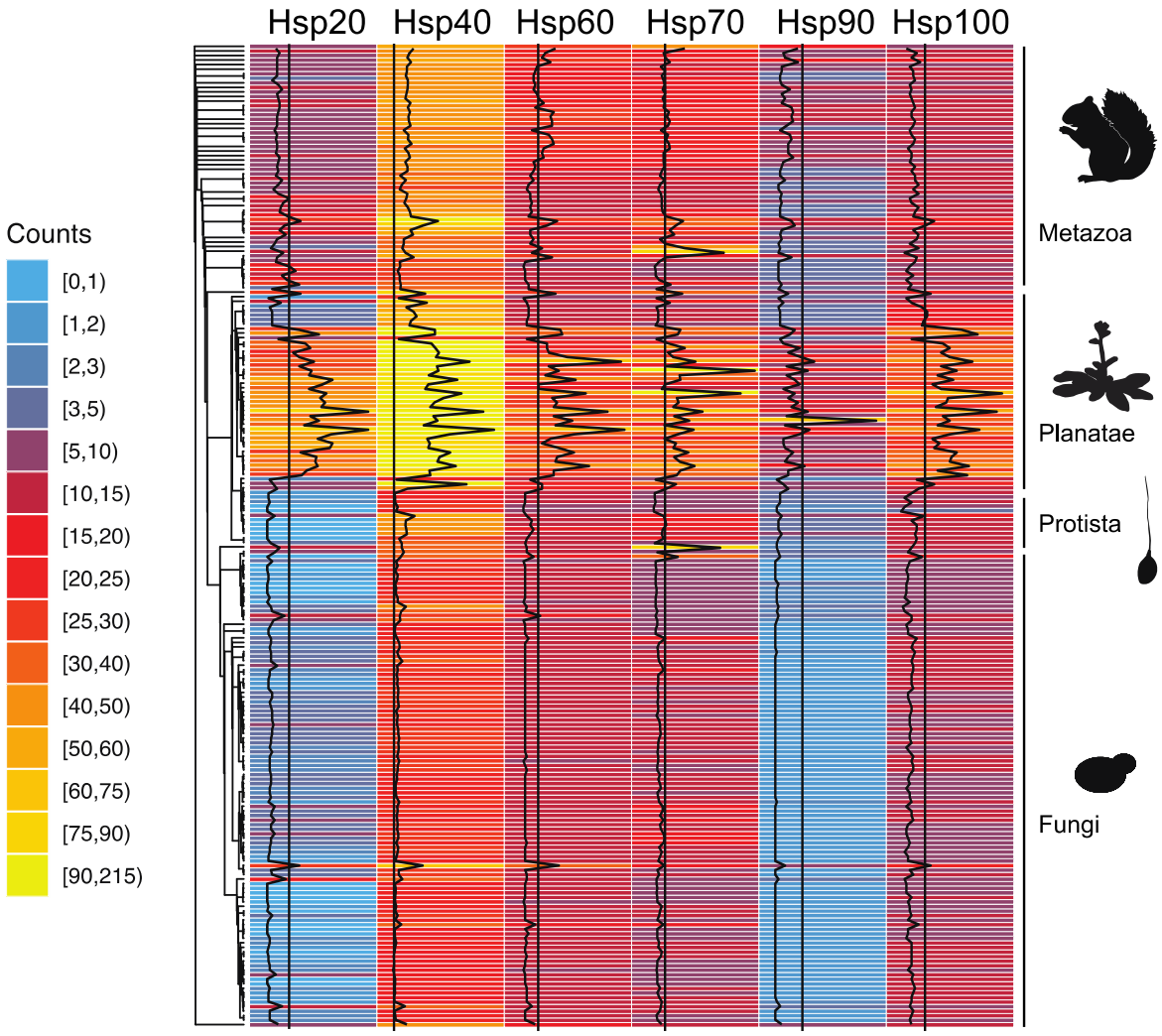

Figure 3.2. Evolution of chaperone networks. The counts of nuclear-encoded Hsp20, Hsp40, Hsp60, Hsp70, Hsp90, and Hsp100 chaperone genes from 216 eukaryotic genomes are visualized as heatmap. Black lines indicate the absolute numbers of chaperone genes relative to their median across all species for each chaperone family.

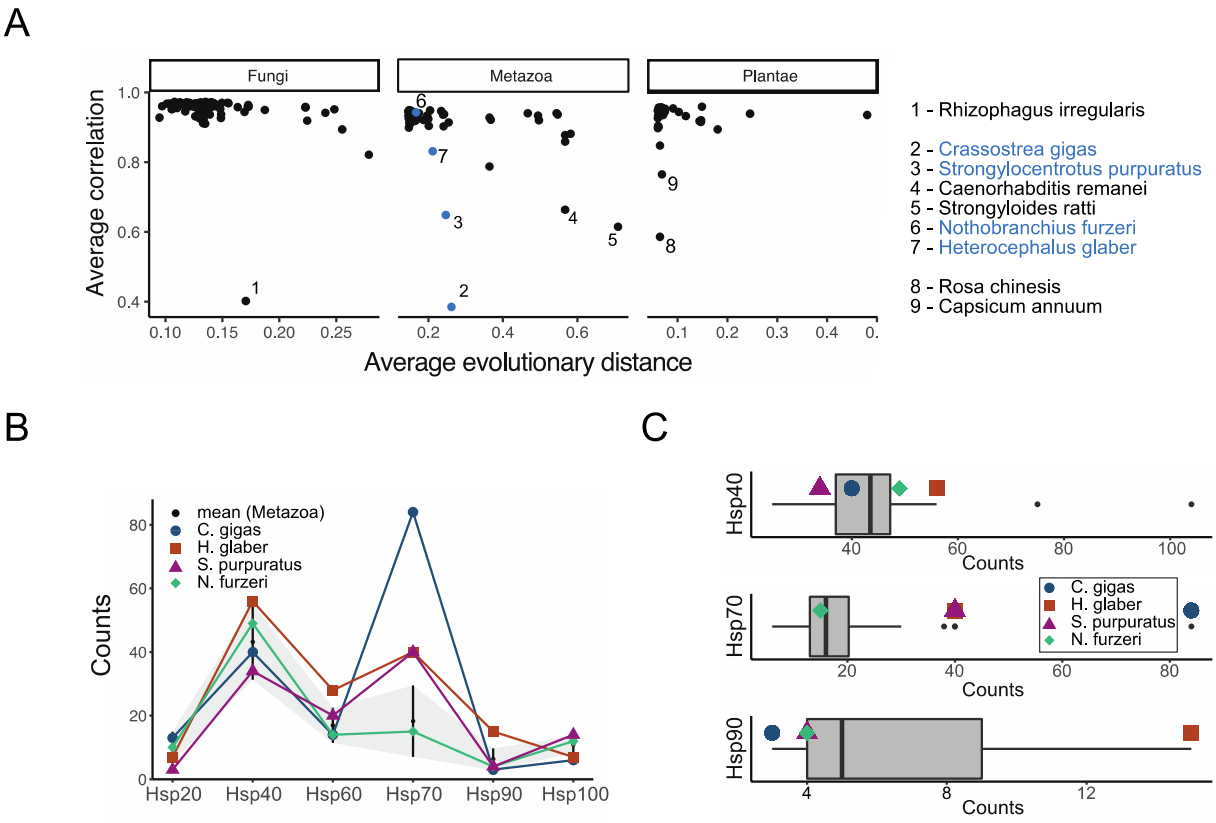

Figure 3.3. Conservation and diversity of in the composition of chaperone networks. **A** Similarity of the composition of chaperone networks within phylogenetic clusters as function of evolutionary distance. Average evolutionary distance denotes the average of pairwise evolutionary distances computed for the branch lengths. Average correlation coefficients represent the average of pairwise correlation coefficients between the composition of chaperone networks of a query species to the other species. **B** Composition of the chaperone networks of four exemplary and noteworthy species that are characterized by known extreme phenotypes, namely the shortest-lived vertebrate *N.furzeri*, the longest lived rodent *H.glaber*, the pacific oyster *C.gigas*, and the sea urchin *S.purpuratus*. As reference are shown the distributions of the chaperone counts in all Metazoa (black dots) +/- standard deviation (grey area). **C** Distribution of the number of chaperones of different families across the animal kingdom. The four species are highlighted in their extreme positions relative to all other animals analyzed.

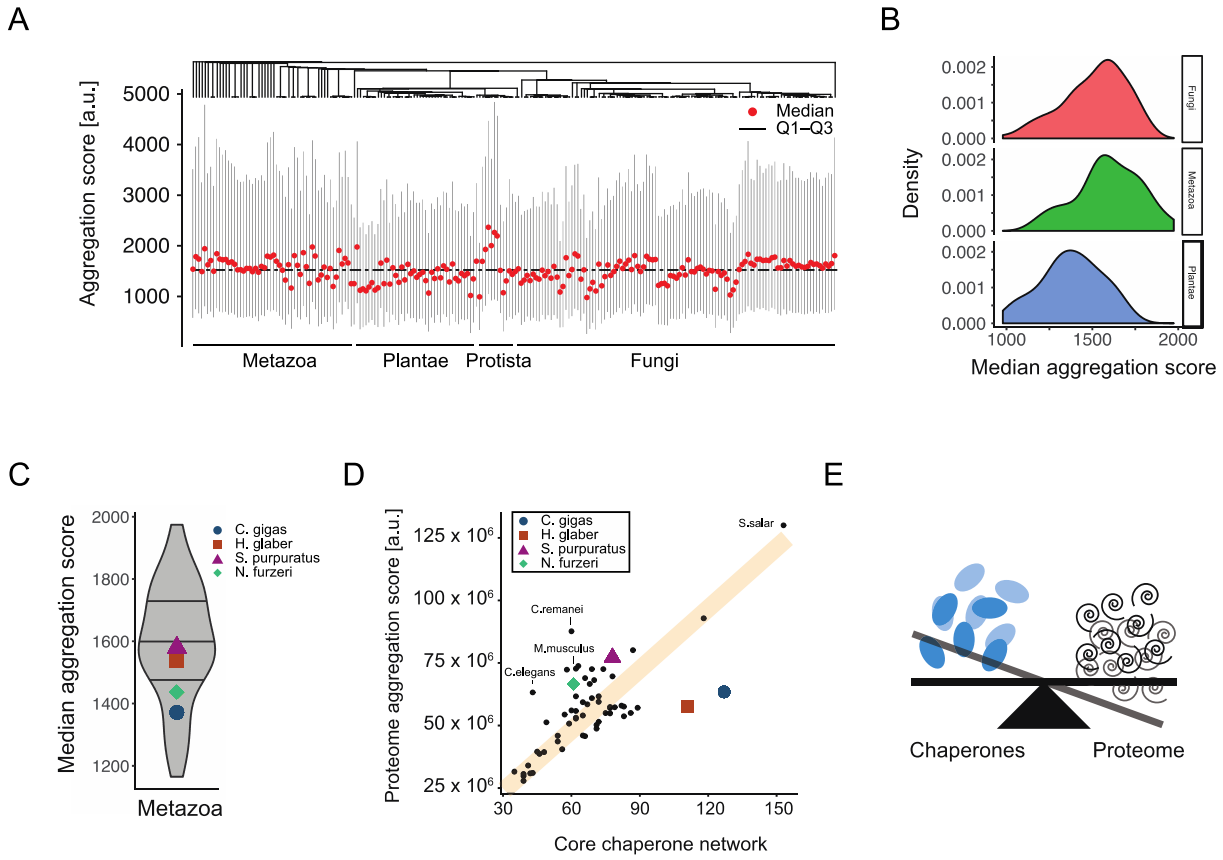

Figure 3.4. Keeping the proteome in balance. A Distribution of predicted protein aggregation propensities in proteomes across the phylogeny. Per-protein predicted aggregation scores are represented by median and 25% (Q1) and 75% (Q3) percentiles for each species. B Distributions of the representative proteome median aggregation scores for the four clusters of species kingdoms on the phylogenetic tree, namely Fungi, Protista, Metazoa, and Plantae. C Distribution of median aggregation scores across the animal kingdom. The median aggregation scores of *C.gigas*, *S. purpuratus*, *N.furzeri* and *H.glaber* are highlighted. D Relationship between the size of chaperone networks and proteome aggregation propensities. The protome aggregation score was computed as the sum of the per-protein predicted aggregation scores, thus reflecting the size of a proteome weighted by the predicted propensity of its proteins to aggregate. The trend line (orange line) is indicative and not a linear fit to the data. E Chaperone networks and proteome co-evolve to maintain balance in protein homeostasis while extreme phenotypes correlate with extreme chaperone counts.

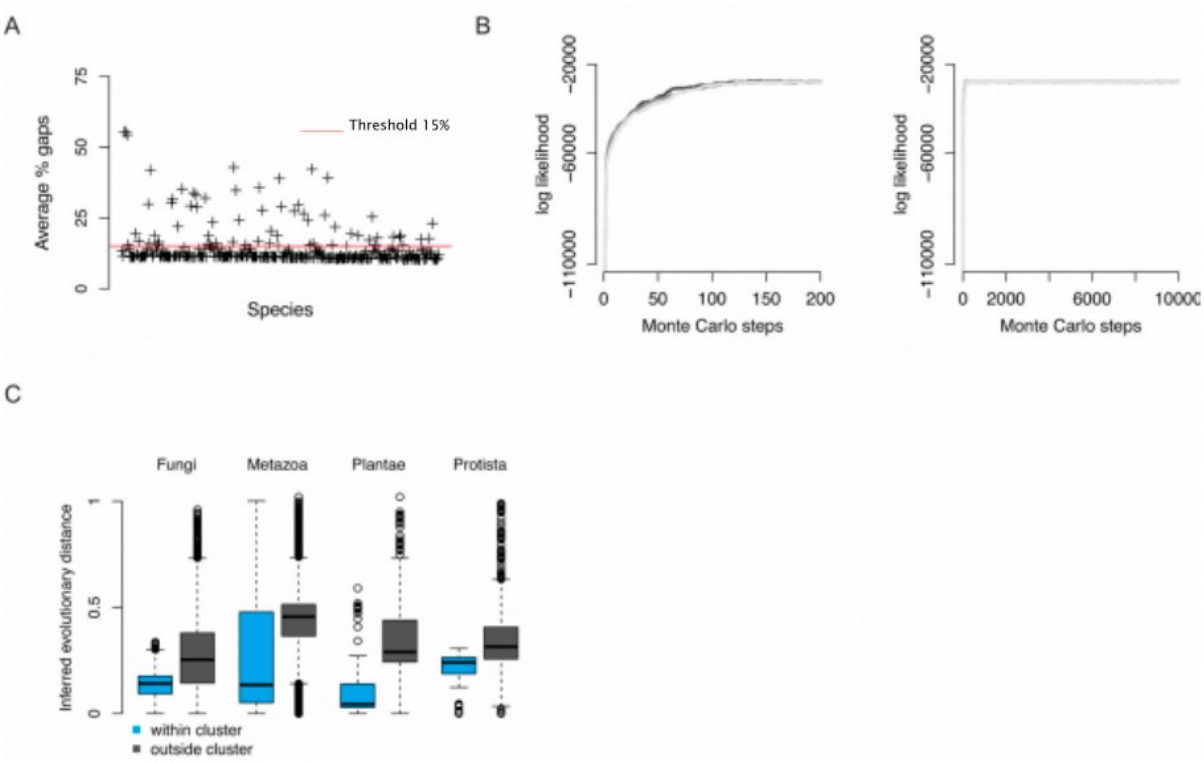

Figure 3.S1. Re-construction and validation of a phylogenetic tree of eukaryotes. A Identification of 18S rRNA sequences that introduce an above average number of gaps into the multiplesequence alignment of eukaryotic 18S rRNA. For each species, the 'average % gaps' score denotes the average number of gaps relative to alignment length in pairwise sequence alignments of a query 18S sequence to the 18S sequences of all other species. The final phylogenetic tree was computed for a dataset without sequences that introduced on average more than 15% gaps (red line). B Convergence of phylogenetic inference. Markov Chain Monte Carlo (MCMC) sampling converges quickly and reproducibly for three independent chains (left: zoom-in on first 200 steps; right: full traces) C Comparison of the distributions of pairwise evolutionary distances inferred from the branch lengths of the phylogenetic tree within and outside of clusters as defined by taxonomy annotations. The observed throughout shorter evolutionary distances between species from the same kingdom as compared to species outside the kingdom underline the quality of the phylogenetic model.

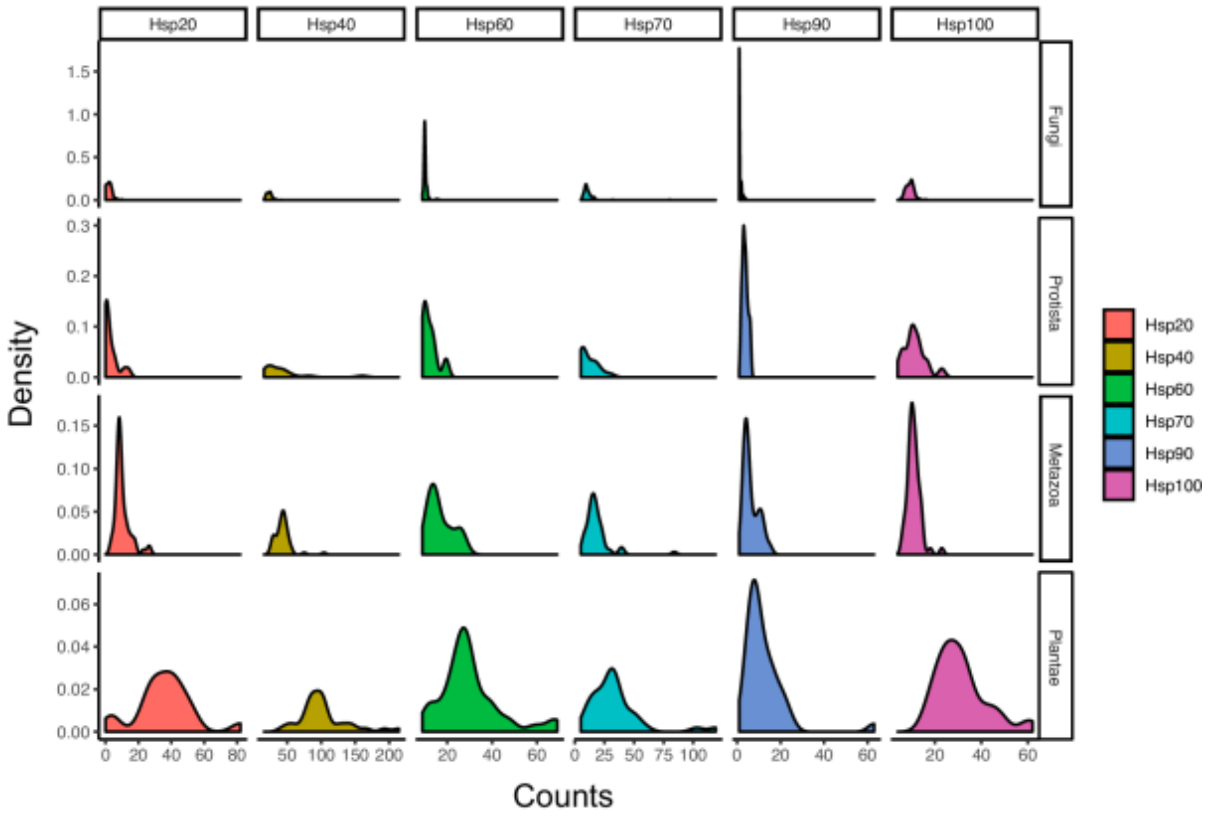

Figure 3.S2. Diversity in chaperone families across the eukaryotic tree of life. Distributions of the numbers of chaperone genes in the chaperone families Hsp20, Hsp40, Hsp60, Hsp70, Hsp90, and Hsp100 are shown for the four kingdoms of Fungi, Protista, Metazoa, and Plantae.

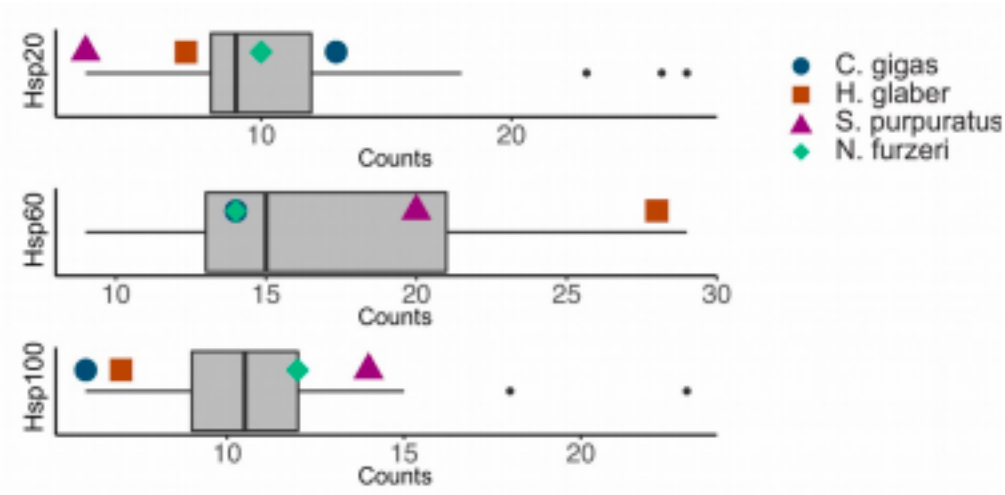

Figure 3.S3. Distribution of the number of Hsp20, Hsp60, and Hsp100 chaperone genes in animal genomes. The determined counts of *C.gigas*, *H.glaber*, *S.purpuratus*, and *N.furzeri* are highlighted.

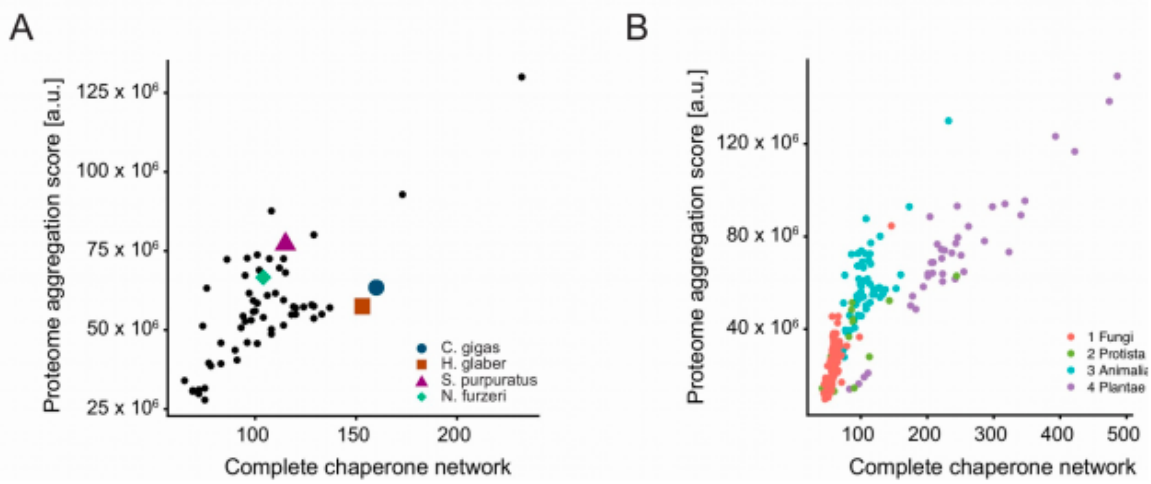

Figure 3.S4. Keeping the proteome in balance. A Relationship between the size of the complete chaperone network and proteome aggregation propensity for animal genomes. B Relationship between the size of the complete chaperone network and proteome aggregation propensity for all studied genomes. The different kingdoms follow distinct trends.

18S rRNA	BLAST	RNAmmer	Our pipeline
Count	88/472	119/472	296/472
Percentage	18.6	25.1	62.4

Table 3.S1. Identification of 18S rRNA sequences in genomes. Shown are the counts of confidently obtained 18S rRNA sequences compared to the number of genomes analyzed by BLAST, RNAmmer, and our new pipeline that combines NGS rRNA read mapping with RNAmmer analysis. The two additional genomes of *N.furzeri* and *H.glaber* were analyzed separately.

Average % gap threshold	Number of sequences	Max diff
15 %	216	0.268
20 %	247	0.540
25 %	253	0.382
30 %	262	0.855
50 %	269	0.440
100 %	270	0.534

Table 3.S2. Benchmarking phylogenetic inference with PhyloBayes. The maxdiff difference between independent MCMC chains is shown for datasets that were filtered at 'average % gaps' thresholds of 15%, 20%, 25%, 30%, 50%, and 100%. A maxdiff < 0.3 is usually considered sufficient for a good model fit, whereas higher values suggest under sampling. The 18S sequences of 28 genomes contained "N" characters and were thus removed from our dataset.

Chapitre 4

Discussion

4.1 Discussion générale

Notre étude in-silico rapporte le nombre de gènes qui codent pour les chaperons, utilisé comme un proxy pour identifier leurs profils d'expression au sein de 216 espèces eucaryotes. Nos résultats appuient les études antérieures (Powers & Balch, 2013; Scheper et al., 2011) sur le fait que les organismes étudiés expriment tous les classes de chaperons. Par contre, 26 champignons dans notre jeu de données n'ont pas d'HSP20 (exemple : *Candida auris*, *Phycomyces blakesleeanus*), ceci pourrait être dû au manque des champignons à la base de données utilisée (HSPIR) lors de la recherche d'homologie des familles de chaperons par HMMER ou des mauvaises annotations (Schnoes, Brown, Dodevski, & Babbitt, 2009) qui induisent des erreurs aux fichiers FASTA sur les données publiques en génomique.

Les espèces qui tolèrent les températures extrêmes froides ou chaudes ont une grande quantité en chaperons. Ces HSP pourraient être le moyen d'adaptation le plus favorable à l'environnement cas *S.purpuratus* pour sa thermotolérance (Hammond & Hofmann, 2010). Il est aussi probable que les espèces les plus sensibles au stress ont des expressions plus élevées aux HSP (Hoekstra, Iwama, Nichols, Godin, & Cheng, 1998; Jolly & Morimoto, 2000). Dans notre jeu de données, les organismes du règne végétale présente le groupe qui a les nombres les plus excessifs de chaperons (Sachs & Ho, 1986; Schulze, Beck, & Hohenstein, 2005). En effet, les plantes ont plusieurs facteurs de stress environnement parmi eux : température extrême, contamination aux métaux lourds, exposition aux radiations ultraviolets, stress

hydrique (Sachs & Ho, 1986), et autre stress induit par des infections aux pathogènes (virales) (Nagy, Wang, Pogany, Hafren, & Makinen, 2011). Ces fortes expressions causées par le stress contribuent au bon maintien de l'homéostasie des protéines par la voie de HSR (Kotak et al., 2007; Timperio, Egidi, & Zolla, 2008). Au niveau génomique, nous expliquons le nombre extrême des six classes des chaperons chez les plantes par leur polyploïdie. Précisément, les plantes polyploïdes peuvent avoir « un niveau d'expression d'un gène polyploïde supérieur ou inférieur au niveau d'expression du gène chez les parents diploïdes » (Combes Gavalda, 2017) ceci définit la catégorie non-additivité (Z. J. Chen, 2010) du niveau d'expression des gènes allopolyploïdes[9]. La surexpression des gènes pourrait être expliqué par des mécanismes génétiques. Parmi-eux, les réarrangements génomiques, cas de duplication des gènes, qui entraine un nombre variable de copie des polyploïdes. Insertion des éléments transposables à côté des gènes qui peuvent conduire à des modifications de l'expression des gènes (cas des régions de méthylation qui conduisent à des modifications transcriptionnelles). Ces mécanismes génétiques sont plus courants sous l'action du stress environnement qui jouent un rôle majeur dans l'adaptation des plantes polyploïdes à des nouveaux environnements (Combes Gavalda, 2017).

L'HSP70 est au cœur de l'expression lors d'un stress pour garder l' homéostasie cellulaire (Deka & Saha, 2018). Nos résultats montrent une variabilité sur le nombre de gènes qui codent pour l'HSP70 corrélé avec le nombre de gènes qui codent pour le co-chaperon HSP40. En littérature, l'expression d'HSP70 dans les mêmes conditions de stress varie entre les types de

[9] Pas issu du même génome.

tissus et cellules (exemple : faible induction d'HSP70 dans les cellules neuronales lors d'un stress) (Deka & Saha, 2018). En condition normal, le taux d'expression des chaperons sur les tissus humains est aussi très variable. Par exemple, il y a une forte expression d'HSP70 sur les cellules neuronales, le tissu lymphoïde des tonsilles, tissus musculaires, épiderme, les tissus du tube digestive et le tissu testiculaire pour l'homme et vaginale pour la femme. Quant à l'HSP40, forte expression sur les tissus endocriniens, les tissus de la vésicule biliaire, épiderme et le tissu testiculaire pour l'homme et les cellules d'ovaires, cellules épithéliales squameuses du col de l'utérus et d'utérine (Berglund et al., 2008; Thul et al., 2017; Uhlén et al., 2015; Uhlen et al., 2017). Ces exemples dévoilent une complexité sur l'expression des chaperons aux niveaux tissulaires et cellulaires. Dans notre cas, nous ne pouvons pas expliquer au-delà sur la variabilité de nos résultats car nous sommes limités par les données disponibles en ligne. Simplement, car à ce jour, il n'y a pas de disposition de données d'expression des HSP au sein des espèces eucaryotiques appart quelques-unes très étudiées comme l'humain. Nos résultats pourront être plus précis s'il y'aurait plus de données d'expression des chaperons pour tous les organismes étudiés.

Pour les animaux, notre jeu de données ressort quelques-uns qui expriment différemment comparant aux autres espèces animales. Nous soulignons : *H.glaber*, *S.purpuratus*, *N.furzeri* et *C.gigas*. Chacune des espèces a un résultat remarquable sur le nombre de chaque famille d'HSP par rapport aux autres organismes. *H.glaber* a le nombre le plus élevé d'HSP90 comparant au *N.furzeri* qui est le plus petit nombre au sein de la même classe. HSP90 intervient dans la réponse au stress cellulaire, dans la signalisation hormonale, dans la survie cellulaire ainsi que le contrôle du cycle cellulaire, celles-ci constituent des processus cellulaires fondamentaux (J. Li, Soroka, & Buchner, 2012). Quant à *C.gigas*, présente le

nombre le plus élevé en HSP70, ceci confirme plusieurs études (Clegg et al., 1998; Hamdoun, Cheney, & Cherr, 2003) sur l'expression de la famille d'HSP70 relié aux stress environnementaux qui accroît encore une fois à la thermotolérance de l'espèce. Nos résultats démontrent aussi, pour la première fois, un nombre plus élevé d'HSP100 pour l'espèce *S.purpuratus*. Aucune étude antérieure n'a élucidé la relation unique entre *S.purpuratus* et HSP100.

En condition normale et en condition de stress, les chaperons stabilisent les protéines afin d'éviter toute forme d'agrégation. Nos résultats de la distribution des propensions à l'agrégation de protéines dans les protéomes montrent une différence qui distingue les quatre groupes. Comparant à l'étude sur 9 espèces eucaryotes (G. G. Tartaglia et al., 2005) qui montre qu'il y a moins de propension à l'agrégation beta sur les séquences des protéomes d'espèces eucaryotiques complexes qui vivent longtemps. Notre étude approuve ce fait, par la comparaison des 216 organismes, les protistes qui sont les moins évolués dans le groupe ont le score le plus élevé de propension à l'agrégation. Cependant, en moyenne, les plantes présentent des scores d'agrégations inférieurs aux champignons et animaux (figure 3.4 A). Le score d'agrégation médian des protéomes de *C.gigas*, *S.purpuratus*, *H.glaber* et *N.fuzeri* se situait en bas du 25% percentile. Ces résultats indiquent que les espèces les plus complexes et évalués ont dans leurs séquences moins de proportion d'agrégation. Véritablement, ce sont les mutations et la sélection évolutive qui ajustent les protéines pour qu'elles soient fonctionnelles. Leurs propensions d'agrégations sont nécessaires à la cellule. Dans le but qu'elles soient assez solubles pour exercer leurs rôles biologiques, les protéines doivent co-évoluer avec leurs environnements cellulaires (Gian Gaetano Tartaglia et al., 2007). Et donc,

le protéome s'adapte aussi à l'environnement et subit une forte pression évolutive que la cellule entraîne pour éviter l'agrégation des protéines (S. Pechmann et al., 2009). Par le fait, une étude in-vivo faite sur l'expression des gènes et le taux d'agrégations des protéines humaines (Gian Gaetano Tartaglia, Pechmann, Dobson, & Vendruscolo, 2007) affirme la coévolution entre le système de chaperons et le protéome. Notre étude in-silico confirme cette coévolution entre les deux aspects génomiques à l'échelle eucaryote. Nous concluons que les chaperons et le protéome co-évoluent pour maintenir l'équilibre de la protéostasie et d'éviter toute forme d'agrégations, tandis que les espèces à phénotypes extrêmes (thermotolérants et résistants au stress) sont en corrélation avec le nombre excessif de chaperons (Figure 3.4 E).

4.2 Méthodologie en phylogénie

L'emplacement des eucaryotes parmi les archées et les bactéries n'est toujours pas résolue selon la dernière représentation de l'arbre (Hug et al., 2016). Il faut dire que les études phylogénétiques sont toujours en cours afin d'élucider des groupes au sein des eucaryotes et parfois des changements majeurs sur les positions des clades au sein de l'arbre. Dans le cas de notre étude, nous avons inférer un arbre de 216 espèces de groupes divergents appartenant au royaume d'eucaryote. Le choix du marqueur phylogénétique et l'utilisation d'un seul marqueur qui évolue lentement sont le plus adéquat pour ces groupes divergents (Patwardhan et al., 2014). Autrement, l'utilisation de plusieurs marqueurs protéiques ainsi que l'application de la méthode de super matrice sont efficaces pour grouper des espèces au sein de la même population (Patwardhan et al., 2014). Le choix du modèle évolutif constitue une étape cruciale. CAT-GTR est probablement le modèle le plus performant pour les données d'ARN et d'ADN et significativement plus robuste contre l'attraction de longues branches de l'arbre, c'est-à-dire, qui regroupent des espèces qui évoluent rapidement sans lien de parenté (Nicolas

Lartillot et al., 2018; Nicolat Lartillot, 2018). Le résultat de l'arbre généré a un écart observé (maxdiff) inférieur 0.3, soit 0.268, qui est acceptable selon l'échelle d'efficacité d'un arbre de consensus (Nicolas Lartillot et al., 2018), nous observons sur la figure 3.1 D que les espèces sont bien regroupés selon la taxonomie. Malgré les bons regroupements, il reste quand même des défis à améliorer, certains ne sont pas bien situé au sein du groupe. Cela peut être dû à la séquence primaire utilisée (mauvais séquençage ou annotations du génome), au niveau du choix du modèle ou nombre élevé d'échantillons prit pour la méthode d'inférence.

4.3 Perspectives

Dans ce mémoire, nous avons présenté une analyse bio-informatique des chaperons qui sont les principaux composant du système d'homéostasie des protéines. Précisément, une analyse de 216 espèces en utilisant une recherche des six classes des chaperons basée sur les profils d'HMM et en calculant le score de propension d'agrégations pour tous les protéomes des organismes étudiés. Ces résultats suggèrent que les chaperons ainsi que les propriétés physico-chimique des protéines sont des facteurs importants de l'homéostasie cellulaire.

Dans le futur, nous aimerions étendre le nombre d'espèces qui sont forts intéressants phénotypiquement parlant. Par exemple, des bactéries thermophiles qui résistent à des fortes températures. Cette démarche nous clarifiera sur l'adaptation des protéines à des fortes températures en comparaison avec les protéines des autres espèces.

Nous aimerions continuer cette étude en exploitant à l'échelle séquentielles. C'est-à-dire établir une étude comparative au niveau des séquences sur la propension d'agrégations. Aussi, exploiter d'autres propriétés physico-chimiques, telle que, la propension de l'activité à prion et la prédiction des protéines intrinsèquement désordonnées. Cette approche nous montrera le détail de l'évolution du protéome à travers la phylogénie, mais aussi, les « patterns » qui pourrait être découverts pour les espèces qui ont un phénotype forts intéressants.

À l'avenir, nous voulions créer un pipeline robuste qui quantifie les gènes qui codent aux chaperons et les classer par rapport à leurs pourcentages d'identités, emplacement au(x) chromosome(s) en utilisant les métadonnées d'NCBI. Cette démarche nous élucidera plus sur l'histoire évolutif du réseau des chaperons et par cela, nous pourrions connaître plus sur les mécanismes génétiques que les chaperons ont établi lors de leur évolution.

Enfin, ces recommandations ajouteront un poids sur nos résultats et sur la compréhension fondamentale et globale de l'homéostasie cellulaire.

Bibliographie

Albanese, V., Yam, A. Y.-W., Baughman, J., Parnot, C., & Frydman, J. (2006). Systems analyses reveal two chaperone networks with distinct functions in eukaryotic cells. *Cell*, *124*(1), 75–88.

Ali, M. M., Roe, S. M., Vaughan, C. K., Meyer, P., Panaretou, B., Piper, P. W., … Pearl, L. H. (2006). Crystal structure of an Hsp90–nucleotide–p23/Sba1 closed chaperone complex. *Nature*, *440*(7087), 1013.

Alouges, F., & Gerin, L. (2018). Un exemple d'optimisation par chaîne de Markov : Le voyageur du commerce. Retrieved from http://gerin.perso.math.cnrs.fr/Enseignements/TP_VoyageurCommerce.pdf

Anders Krogh. (1998). *An Introduction to Hidden Markov Models for Biological Sequences*. Technical University of Denmark, Lyngby, Denmark.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … Magrane, M. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *32*(suppl_1), D115–D119.

Bakthisaran, R., Tangirala, R., & Rao, C. M. (2015). Small heat shock proteins: role in cellular functions and pathology. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1854*(4), 291–319.

Baumgart, M., Groth, M., Priebe, S., Savino, A., Testa, G., Dix, A., … Ori, M. (2014). RNA-seq of the aging brain in the short- lived fish N. furzeri–conserved pathways and novel genes associated with neurogenesis. *Aging Cell*, *13*(6), 965–974.

Béatrice Roure. (2011). *Amélioration de l'exactitude de l'inférence phylogénomique*. Université de Montréal, Montréal.

Beerten, J., Jonckheere, W., Rudyak, S., Xu, J., Wilkinson, H., De Smet, F., … Rousseau, F. (2012). Aggregation gatekeepers modulate protein homeostasis of aggregating sequences and affect bacterial fitness. *Protein Eng Des Sel*, *25*(7), 357–366. https://doi.org/10.1093/protein/gzs031

Berglund, L., Björling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C. A.-K., …

Uhlén, M. (2008). A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Molecular & Cellular Proteomics*, *7*(10), 2019–2027. https://doi.org/10.1074/mcp.R800013-MCP200

Bershtein, S., Mu, W., Serohijos, A. W. R., Zhou, J., & Shakhnovich, E. I. (2013). Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Molecular Cell*, *49*(1), 133–144. https://doi.org/10.1016/j.molcel.2012.11.004

Bogumil, D., Alvarez-Ponce, D., Landan, G., McInerney, J. O., & Dagan, T. (2014). Integration of two ancestral chaperone systems into one: the evolution of eukaryotic molecular chaperones in light of eukaryogenesis. *Molecular Biology and Evolution*, *31*(2), 410–418. https://doi.org/10.1093/molbev/mst212

Bogumil, David, Landan, G., Ilhan, J., & Dagan, T. (2012). Chaperones divide yeast proteins into classes of expression level and evolutionary rate. *Genome Biology and Evolution*, *4*(5), 618–625. https://doi.org/10.1093/gbe/evs025

Bosco, D. A., LaVoie, M. J., Petsko, G. A., & Ringe, D. (2011). Proteostasis and Movement Disorders: Parkinson's Disease and Amyotrophic Lateral Sclerosis. *Cold Spring Harbor Perspectives in Biology*, *3*(10). https://doi.org/10.1101/cshperspect.a007500

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, *6*(5), a016147.

Caetano-Anolles, D., Kim, K. M., Mittenthal, J. E., & Caetano-Anolles, G. (2011). Proteome evolution and the metabolic origins of translation and cellular life. *Journal of Molecular Evolution*, *72*(1), 14–33. https://doi.org/10.1007/s00239-010-9400-9

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Cameron, R. A., Samanta, M., Yuan, A., He, D., & Davidson, E. (2009). SpBase: the sea urchin genome database and web site. *Nucleic Acids Research*, *37*(Database issue), D750-754. https://doi.org/10.1093/nar/gkn887

Cannon, W. B. (1932). The wisdom of the body.

Chen, B., Zhong, D., & Monteiro, A. (2006). Comparative genomics and evolution of the HSP90 family of genes across all kingdoms of organisms. *BMC Genomics*, *7*(1), 156.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, *15*(2), 57. https://doi.org/10.1016/j.tplants.2009.12.003

Chiti, F., & Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, *75*, 333–366.

Choo, K. H., Tong, J. C., & Zhang, L. (2004). Recent applications of hidden Markov models in computational biology. *Genomics, Proteomics & Bioinformatics*, *2*(2), 84–96.

Clegg, J. S., Uhlinger, K. R., Jackson, S. A., Cherr, G. N., Rifkin, E., & Friedman, C. S. (1998). Induced thermotolerance and the heat shock protein-70 family in the Pacific oyster Crassostrea gigas. *Molecular Marine Biology and Biotechnology*, *7*, 21–30.

Combes Gavalda, M.-C. (2017). Polyploïdie et adaptation des plantes Étude de l'expression des gènes homéologues chez le caféier (Coffea arabica), 25.

Damien Aubert. (2017). *Classer le vivant: Les perspectives de la systématique évolutionniste moderne* (ELLIPSES).

Darlu, P., & Tassy, P. (1993). La reconstruction phylogénétique. *Concepts et Méthodes. Paris, France*.

David, D. C. (2012). Aging and the aggregating proteome. *Front Genet*, *3*, 247. https://doi.org/10.3389/fgene.2012.00247

Davies, P. L. (2014). Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. *Trends Biochem Sci*, *39*(11), 548–555. https://doi.org/10.1016/j.tibs.2014.09.005

Dayrat, B. (2003). The roots of phylogeny: how did Haeckel build his trees? *Systematic Biology*, *52*(4), 515–527.

de Queiroz, A., & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology & Evolution*, *22*(1), 34–41.

Deka, K., & Saha, S. (2018). Regulation of Mammalian HSP70 Expression and Stress Response. In A. A. A. Asea & P. Kaur (Eds.), *Regulation of Heat Shock Protein Responses* (pp. 3–25). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-74715-6_1

Dikic, I. (2017). Proteasomal and Autophagic Degradation Systems. *Annual Review of Biochemistry*, *86*, 193–224. https://doi.org/10.1146/annurev-biochem-061516-044908

Dobson, C. M., & Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends in Biochemical Sciences*, *24*(9), 329–332. https://doi.org/10.1016/S0968-0004(99)01445-0

Drost, H.-G., Gabel, A., Liu, J., Quint, M., & Grosse, I. (2017). myTAI: evolutionary

transcriptomics with R. *Bioinformatics*, *34*(9), 1589–1590.
Eddy, S. (1992). HMMER user's guide. *Department of Genetics, Washington University School of Medicine*, *2*(1).
Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.
Edwards, H. V., Scott, J. D., & Baillie, G. S. (2012). PKA phosphorylation of the small heat-shock protein Hsp20 enhances its cardioprotective effects. Portland Press Limited.
Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, *8*(3), 163–167.
Ellis, R. J. (1996). Discovery of molecular chaperones. *Cell Stress & Chaperones*, *1*(3), 155.
Eric Paul Nawrocki. (2009). *STRUCTURAL RNA HOMOLOGY SEARCH AND ALIGNMENT USING COVARIANCE MODELS*. WASHINGTON UNIVERSITY, Washington.
Evangella Kranias. (2015). Heat Shock Protein 20.
Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, *22*(10), 1302.
Frédéric Delsuc. ; Emmanuel Douzery. (2004). Les méthodes probabilistes en phylogénie moléculaire: (2) L'approche bayésienne. *Biosystema, Société Française de Systématique*, *22*, 75–86.
Frydman, J. (2001). Folding of newly translated proteins in vivo: the role of molecular chaperones. *Annual Review of Biochemistry*, *70*(1), 603–647.
Fulda, S., Gorman, A. M., Hori, O., & Samali, A. (2010). Cellular stress responses: cell survival and cell death. *Int J Cell Biol*, *2010*(2010), 214074.
futura planète. (2018). phylogénie.
Geller, R. (2007). Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes & Development*, *21(2)*, 195–205.
Geller, Ron, Pechmann, S., Acevedo, A., Andino, R., & Frydman, J. (2018). Hsp90 shapes protein and RNA evolution to balance trade-offs between protein stability and aggregation. *Nature Communications*, *9*(1), 1781. https://doi.org/10.1038/s41467-018-04203-x
Goldstone, J. V., Hamdoun, A., Cole, B. J., Howard-Ashby, M., Nebert, D. W., Scally, M., …

Stegeman, J. J. (2006). The chemical defensome: environmental sensing and response genes in the Strongylocentrotus purpuratus genome. *Developmental Biology*, *300*(1), 366–384.
Greaves, R. B., & Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC Struct Biol*, *7*, 18. https://doi.org/10.1186/1472-6807-7-18
Greene, M. K., Maskos, K., & Landry, S. J. (1998). Role of the J-domain in the cooperation of Hsp40 with Hsp70. *Proceedings of the National Academy of Sciences*, *95*(11), 6108–6113.
Hadizadeh Esfahani, A., Sverchkova, A., Saez-Rodriguez, J., Schuppert, A. A., & Brehme, M. (2018). A systematic atlas of chaperome deregulation topologies across the human cancer landscape. *PLoS Computational Biology*, *14*(1), e1005890. https://doi.org/10.1371/journal.pcbi.1005890
Hamdoun, A. M., Cheney, D. P., & Cherr, G. N. (2003). Phenotypic plasticity of HSP70 and HSP70 gene expression in the Pacific oyster (Crassostrea gigas): implications for thermal limits and induction of thermal tolerance. *The Biological Bulletin*, *205*(2), 160–169.
Hammond, L. M., & Hofmann, G. E. (2010). Thermal tolerance of Strongylocentrotus purpuratus early life history stages: mortality, stress-induced gene expression and biogeographic patterns. *Marine Biology*, *157*(12), 2677–2687.
Harel, I., Valenzano, D. R., & Brunet, A. (2016). Efficient genome engineering approaches for the short-lived African turquoise killifish. *Nature Protocols*, *11*(10), 2010.
Hartl, F. U. (1996). Molecular chaperones in cellular protein folding. *Nature*, *381*(6583), 571.
Hartl, F. U., & Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, *295*(5561), 1852–1858.
Hillis, D. M., & Dixon, M. T. (1991). Ribosomal DNA: Molecular Evolution and Phylogenetic Inference. *The Quarterly Review of Biology*, *66*(4), 411–453.
Hoekstra, K. A., Iwama, G. K., Nichols, C. R., Godin, D. V., & Cheng, K. M. (1998). Increased heat shock protein expression after stress in Japanese quail. *Stress (Amsterdam, Netherlands)*, *2*(4), 265–272.
Horan, K., Shelton, C. R., & Girke, T. (2010). Predicting conserved protein motifs with Sub-HMMs. *BMC Bioinformatics*, *11*(1), 205.
Huelsenbeck, J. P., & Rannala, B. (1997). Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context. *Science*, *276*(5310), 227–232.

https://doi.org/10.1126/science.276.5310.227
Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, *294*(5550), 2310–2314.
Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., … Ise, K. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 16048.
Imai, J., Maruya, M., Yashiroda, H., Yahara, I., & Tanaka, K. (2003). The molecular chaperone Hsp90 plays a role in the assembly and maintenance of the 26S proteasome. *The EMBO Journal*, *22*(14), 3557–3567.
Jia, F., Lo, N., & Ho, S. Y. (2014). The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. *PLoS One*, *9*(5), e95722.
Jolly, C., & Morimoto, R. I. (2000). Role of the Heat Shock Response and Molecular Chaperones in Oncogenesis and Cell Death. *JNCI: Journal of the National Cancer Institute*, *92*(19), 1564–1572. https://doi.org/10.1093/jnci/92.19.1564
Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, *3*(21), 132.
Jürgen Radons. (2018a). HSP60: FIGURES.
Jürgen Radons. (2018b). HSP60: FUNCTION.
Jürgen Radons. (2018c). HSP90: STRUCTURE.
Kaushik, S., & Cuervo, A. M. (2015). Proteostasis and aging. *Nature Medicine*, *21*(12), 1406.
Keane, M., Craig, T., Alfoldi, J., Berlin, A. M., Johnson, J., Seluanov, A., … de Magalhaes, J. P. (2014). The Naked Mole Rat Genome Resource: facilitating analyses of cancer and longevity-related adaptations. *Bioinformatics (Oxford, England)*, *30*(24), 3558–3560. https://doi.org/10.1093/bioinformatics/btu579
Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357.
Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M., & Ulrich Hartl, F. (2013). Molecular chaperone functions in protein folding and proteostasis. *Annual Review of Biochemistry*, *82*, 323–355.
Kim, Y., Nam, H. G., & Valenzano, D. R. (2016). The short-lived African turquoise killifish: an emerging experimental model for ageing. *Disease Models & Mechanisms*, *9*(2), 115–129.

Kimberly Riskas. (2018). For naked mole-rats, breeding is key to longevity.
Klaips, C. L., Jayaraj, G. G., & Hartl, F. U. (2018). Pathways of cellular proteostasis in aging and disease. *The Journal of Cell Biology*, *217*(1), 51. https://doi.org/10.1083/jcb.201709072
Koga, H., Kaushik, S., & Cuervo, A. M. (2011). Protein homeostasis and aging: The importance of exquisite quality control. *Ageing Research Reviews*, *10*(2), 205–215.
Korcsmaros, T., Kovacs, I. A., Szalay, M. S., & Csermely, P. (2007). Molecular chaperones: the modular evolution of cellular networks. *J Biosci*, *32*(3), 441–446.
Kotak, S., Larkindale, J., Lee, U., von Koskull-Döring, P., Vierling, E., & Scharf, K.-D. (2007). Complexity of the heat stress response in plants. *Current Opinion in Plant Biology*, *10*(3), 310–316.
Kultz, D. (2003). Evolution of the cellular stress proteome: from monophyletic origin to ubiquitous function. *J Exp Biol*, *206*(Pt 18), 3119–3124.
Labbadia, J., & Morimoto, R. I. (2015). The biology of proteostasis in aging and disease. *Annual Review of Biochemistry*, *84*, 435–464.
Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108.
Lartillot, Nicolas, Blanquart, S., & Lepage, T. (2018). Phylobayes 3.3 a Bayesian software for phylogenetic reconstruction and molecular dating using mixture models. Retrieved from http://megasun.bch.umontreal.ca/People/lartillot/www/phylobayes3.3e.pdf
Lartillot, Nicolas, Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, *25*(17), 2286–2288.
Lartillot, Nicolas, & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109.
Lartillot, Nicolas, Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, *62*(4), 611–615.
Lartillot, Nicolat. (2018). Choosing the right model.
Leach, M. D., Klipp, E., Cowen, L. E., & Brown, A. J. (2012). Fungal Hsp90: a biological

transistor that tunes cellular outputs to thermal inputs. *Nature Reviews Microbiology*, *10*(10), 693.

Leitner, A., Joachimiak, L. A., Bracher, A., Mönkemeyer, L., Walzthoeni, T., Chen, B., … Ma, B. (2012). The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure*, *20*(5), 814–825.

Li, J., Soroka, J., & Buchner, J. (2012). The Hsp90 chaperone machinery: conformational dynamics and regulation by co-chaperones. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1823*(3), 624–635.

Li, S., Pearl, D. K., & Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, *95*(450), 493–508.

Lim, H.-J., Kim, B.-M., Hwang, I. J., Lee, J.-S., Choi, I.-Y., Kim, Y.-J., & Rhee, J.-S. (2016). Thermal stress induces a distinct transcriptome profile in the Pacific oyster Crassostrea gigas. *Comparative Biochemistry and Physiology. Part D, Genomics & Proteomics*, *19*, 62–70. https://doi.org/10.1016/j.cbd.2016.06.006

Lindquist, S., & Craig, E. A. (1988). The heat-shock proteins. *Annual Review of Genetics*, *22*(1), 631–677.

Luis Serrano ; Joost Schymkowitz ; Frederic Rousseau. (2018). *Tango Handbook*.

M. Mariadassou ; A. Bar-Hen. (2010). *Introduction `a la Phylog´enie mol´eculaire*. Paris.

Malinovska, L., Palm, S., Gibson, K., Verbavatz, J. M., & Alberti, S. (2015). Dictyostelium discoideum has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation. *Proc Natl Acad Sci U S A*, *112*(20), E2620-9. https://doi.org/10.1073/pnas.1504459112

Mau, B., Newton, M. A., & Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, *55*(1), 1–12.

McClellan, A. J., Xia, Y., Deutschbauer, A. M., Davis, R. W., Gerstein, M., & Frydman, J. (2007). Diverse cellular functions of the Hsp90 molecular chaperone uncovered using systems approaches. *Cell*, *131*(1), 121–135.

Milisav, I. (2011). Cellular stress responses. In *Advances in regenerative medicine*. InTech.

Monsellier, E., Ramazzotti, M., Taddei, N., & Chiti, F. (2008). Aggregation propensity of the human proteome. *PLoS Computational Biology*, *4*(10), e1000199. https://doi.org/10.1371/journal.pcbi.1000199

Musi, N., & Hornsby, P. (2015). *Handbook of the Biology of Aging*. Academic Press.
Myers, P. (2018). The Animal Diversity Web (online).
Nagy, P. D., Wang, R. Y., Pogany, J., Hafren, A., & Makinen, K. (2011). Emerging picture of host chaperone and cyclophilin roles in RNA virus replication. *Virology*, *411*(2), 374–382. https://doi.org/10.1016/j.virol.2010.12.061
Nawrocki, E. P., & Eddy, S. R. (2010). ssu-align: a tool for structural alignment of SSU rRNA sequences. URL h ttp://selab. janelia. org/software. html.
Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, *25*(10), 1335–1337.
Nicolaou, P., Knöll, R., Haghighi, K., Fan, G.-C., Dorn, G. W., Hasenfuß, G., & Kranias, E. G. (2008). A human mutation in the anti-apoptotic heat shock protein 20 abrogates its cardioprotective effects. *Journal of Biological Chemistry*.
Nillegoda, N. B., & Bukau, B. (2015). Metazoan Hsp70-based protein disaggregases: emergence and mechanisms. *Frontiers in Molecular Biosciences*, *2*, 57.
Noller, H. F. (1984). Structure of Ribosomal Rna. *Annual Review of Biochemistry*, *53*(1), 119–162. https://doi.org/10.1146/annurev.bi.53.070184.001003
O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Ako-Adjei, D. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745.
Page, R. D., & Holmes, E. C. (2009). *Molecular evolution: a phylogenetic approach*. John Wiley & Sons.
Patrick D'Silva. (2014). Hsp100: AAA+ Family of Chaperones.
Patwardhan, A., Ray, S., & Roy, A. (2014). Molecular markers in phylogenetic studies-a review. *Journal of Phylogenetics & Evolutionary Biology*, *2014*.
Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1. 1-3.1. 8.
Pechmann, S., & Frydman, J. (2014). Interplay between chaperones and protein disorder promotes the evolution of protein networks. *PLoS Computational Biology*, *10*(6), e1003674. https://doi.org/10.1371/journal.pcbi.1003674
Pechmann, S., Levy, E. D., Tartaglia, G. G., & Vendruscolo, M. (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of

proteins. *Proc Natl Acad Sci U S A*, *106*(25), 10159–10164. https://doi.org/10.1073/pnas.0812414106

Pechmann, Sebastian, & Vendruscolo, M. (2010). Derivation of a solubility condition for proteins from an analysis of the competition between folding and aggregation. *Molecular BioSystems*, *6*(12), 2490–2497. https://doi.org/10.1039/c005160h

Pechmann, Sebastian, Willmund, F., & Frydman, J. (2013). The ribosome as a hub for protein quality control. *Molecular Cell*, *49*(3), 411–421.

Picard, D., Khursheed, B., Garabedian, M. J., Fortin, M. G., Lindquist, S., & Yamamoto, K. R. (1990). Reduced levels of hsp90 compromise steroid receptor action in vivo. *Nature*, *348*(6297), 166.

Platzer, M., & Englert, C. (2016). Nothobranchius furzeri: a model for aging research and more. *Trends in Genetics*, *32*(9), 543–552.

Powers, E. T., & Balch, W. E. (2013). Diversity in the origins of proteostasis networks—a driver for protein function in evolution. *Nature Reviews Molecular Cell Biology*, *14*(4), 237.

Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., & Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annual Review of Biochemistry*, *78*, 959–991.

Pratt, W. B., Morishima, Y., Murphy, M., & Harrell, M. (2006). Chaperoning of glucocorticoid receptors. In *Molecular Chaperones in Health and Disease* (pp. 111–138). Springer.

Press, M. O., Li, H., Creanza, N., Kramer, G., Queitsch, C., Sourjik, V., & Borenstein, E. (2013). Genome-scale Co-evolutionary Inference Identifies Functions and Clients of Bacterial Hsp90. *PLOS Genetics*, *9*(7), e1003631. https://doi.org/10.1371/journal.pgen.1003631

Qiu, X.-B., Shao, Y.-M., Miao, S., & Wang, L. (2006). The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cellular and Molecular Life Sciences CMLS*, *63*(22), 2560–2570.

Rannala, B., & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, *43*(3), 304–311.

Reichwald, K., Petzold, A., Koch, P., Downie, B. R., Hartmann, N., Pietsch, S., … Bens, M. (2015). Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell*, *163*(6), 1527–1538.

Reynaud, E. (2010). Protein misfolding and degenerative diseases. *Nature Education*, *3*(9), 28.
Richter, K., Haslbeck, M., & Buchner, J. (2010). The heat shock response: life on the verge of death. *Molecular Cell*, *40*(2), 253–266.
Rodriguez, K. A., Valentine, J. M., Kramer, D. A., Gelfond, J. A., Kristan, D. M., Nevo, E., & Buffenstein, R. (2016). Determinants of rodent longevity in the chaperone-protein degradation network. *Cell Stress and Chaperones*, *21*(3), 453–466.
Ross, C. A., & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, *10*(7), S10.
Roth, D. M. (2011). Modeling general proteostasis: proteome balance in health and disease. *Current Opinion in Cell Biology*, *23(2)*, 126–134.
Rudan, M., Schneider, D., Warnecke, T., & Krisko, A. (n.d.). RNA chaperones buffer deleterious mutations in E. coli. *ELife*, *4*. https://doi.org/10.7554/eLife.04745
Sachs, M. M., & Ho, T. H. D. (1986). Alteration of Gene Expression During Environmental Stress in Plants. *Annual Review of Plant Physiology*, *37*(1), 363–376. https://doi.org/10.1146/annurev.pp.37.060186.002051
Saibil, H. (2000). Molecular chaperones: containers and surfaces for folding, stabilising or unfolding proteins. *Current Opinion in Structural Biology*, *10*(2), 251–258.
Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.
Sala, A. J., Bott, L. C., & Morimoto, R. I. (2017). Shaping proteostasis at the cellular, tissue, and organismal level. *J Cell Biol*, *216*(5), 1231–1241.
Santra, M., Farrell, D. W., & Dill, K. A. (2017). Bacterial proteostasis balances energy and chaperone utilization efficiently. *Proceedings of the National Academy of Sciences*, *114*(13), E2654–E2661.
Scheper, W., Nijholt, D. A., & Hoozemans, J. J. (2011). The unfolded protein response and proteostasis in Alzheimer disease. *Autophagy*, *7*(8), 910–911. https://doi.org/10.4161/auto.7.8.15761
Schneider, K., & Bertolotti, A. (2015). Surviving protein quality control catastrophes--from cells to organisms. *J Cell Sci*, *128*(21), 3861–3869. https://doi.org/10.1242/jcs.173047
Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS*

Computational Biology, *5*(12), e1000605.
Schulze, E. D., Beck, E., & Hohenstein, K. M. (2005). Environment as stress factor: stress physiology of plants. *Plant Ecology*, *702*.
Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., … Wright, R. (2006). The genome of the sea urchin Strongylocentrotus purpuratus. *Science (New York, N.Y.)*, *314*(5801), 941–952. https://doi.org/10.1126/science.1133609
Sinha, D., Veedin Rajan, V. B., Esthaki, V. K., & D'Silva, P. (2012). HSPIR: a manually annotated heat shock protein information resource. *Bioinformatics*, *28*(21), 2853–2855.
Smith, L. C. (2012). Innate immune complexity in the purple sea urchin: diversity of the sp185/333 system. *Frontiers in Immunology*, *3*, 70. https://doi.org/10.3389/fimmu.2012.00070
Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, *21*(7), 951–960. https://doi.org/10.1093/bioinformatics/bti125
Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, *26*(1), 320–322.
Sørensen, J. G. (2010). Application of heat shock protein expression for detecting natural adaptation and exposure to stress in natural populations. *Current Zoology*, *56*(6), 11.
Sőti, C., Pál, C., Papp, B., & Csermely, P. (2005). Molecular chaperones as regulatory elements of cellular networks. *Current Opinion in Cell Biology*, *17*(2), 210–215.
Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & Doyle, J. (2004). Robustness of cellular functions. *Cell*, *118*(6), 675–685. https://doi.org/10.1016/j.cell.2004.09.008
Sun, Y., & MacRae, T. H. (2005). Small heat shock proteins: molecular structure and chaperone function. *Cellular and Molecular Life Sciences CMLS*, *62*(21), 2460–2476.
Taguwa, S., Maringer, K., Li, X., Bernal-Rubio, D., Rauch, J. N., Gestwicki, J. E., … Frydman, J. (2015). Defining Hsp70 Subnetworks in Dengue Virus Replication Reveals Key Vulnerability in Flavivirus Infection. *Cell*, *163*(5), 1108–1123. https://doi.org/10.1016/j.cell.2015.10.046

Taipale, M., Jarosz, D. F., & Lindquist, S. (2010). HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular Cell Biology*, *11*(7), 515.

Tartaglia, G. G., Pellarin, R., Cavalli, A., & Caflisch, A. (2005). Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein Sci*, *14*(10), 2735–2740. https://doi.org/10.1110/ps.051473805

Tartaglia, Gian Gaetano, Dobson, C. M., Hartl, F. U., & Vendruscolo, M. (2010). Physicochemical determinants of chaperone requirements. *Journal of Molecular Biology*, *400*(3), 579–588.

Tartaglia, Gian Gaetano, Pechmann, S., Dobson, C. M., & Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in Biochemical Sciences*, *32*(5), 204–206. https://doi.org/10.1016/j.tibs.2007.03.005

Tartaglia, Gian Gaetano, & Vendruscolo, M. (2008). The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews*, *37*(7), 1395–1401.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*(2), 57–86.

Terzibasi, E., Valenzano, D. R., & Cellerino, A. (2007). The short-lived fish Nothobranchius furzeri as a new model system for aging studies. *Experimental Gerontology*, *42*(1–2), 81–89.

Thangakani, A. M., Kumar, S., Velmurugan, D., & Gromiha, M. S. (2012). How do thermophilic proteins resist aggregation? *Proteins*, *80*(4), 1003–1015. https://doi.org/10.1002/prot.24002

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Blal, H. A., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, *356*(6340), eaal3321. https://doi.org/10.1126/science.aal3321

Timperio, A. M., Egidi, M. G., & Zolla, L. (2008). Proteomics applied on plant abiotic stresses: role of heat shock proteins (HSP). *Journal of Proteomics*, *71*(4), 391–411.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., … Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, *357*(6352), eaan2507. https://doi.org/10.1126/science.aan2507

Uniprot. (2018). UniProtKB - P0DMV8 (HS71A_HUMAN).

Valenzano, D. R., Benayoun, B. A., Singh, P. P., Zhang, E., Etter, P. D., Hu, C. K., … Brunet, A. (2015). The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. *Cell*, *163*(6), 1539–1554. https://doi.org/10.1016/j.cell.2015.11.008

Van de Peer, Y., & De Wachter, R. (1997). Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *Journal of Molecular Evolution*, *45*(6), 619–630.

Wallace, E. W., Kear-Scott, J. L., Pilipenko, E. V., Schwartz, M. H., Laskowski, P. R., Rojek, A. E., … Drummond, D. A. (2015). Reversible, Specific, Active Aggregates of Endogenous Proteins Assemble upon Heat Stress. *Cell*, *162*(6), 1286–1298. https://doi.org/10.1016/j.cell.2015.08.041

Wang, L., Zhang, H., Wang, M., Zhou, Z., Wang, W., Liu, R., … Song, L. (2019). The transcriptomic expression of pattern recognition receptors: Insight into molecular recognition of various invading pathogens in Oyster Crassostrea gigas. *Developmental & Comparative Immunology*, *91*, 1–7.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Willmund, F., del Alamo, M., Pechmann, S., Chen, T., Albanèse, V., Dammer, E. B., … Frydman, J. (2013). The cotranslational function of ribosome-associated Hsp70 in eukaryotic protein homeostasis. *Cell*, *152*(1), 196–209.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, *87*(12), 4576–4579.

Wu, M., & Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, *9*(10), R151.

Yam, A. Y., Xia, Y., Lin, H.-T. J., Burlingame, A., Gerstein, M., & Frydman, J. (2008). Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nature Structural & Molecular Biology*, *15*(12), 1255.

Yan Li. (2018). How to build a phylogenetic tree.

Yu, G., Lam, T. T.-Y., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, *35*(12), 3041–3043.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., … Wang, J. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49–54. https://doi.org/10.1038/nature11413

Annexe

Tableau 2.S1. Distribution des chaperons moléculaires à travers les espèces. Tiré de (Powers & Balch, 2013)

Organism	Type	Genome size (Mb)[a]	Number of Genes[a]	Hsp70[b]	Hsp40[b]	Hsp90[b]	Hsp60 subunits (cytoplasm/ organelles)[b]	sHsp[b]	Hsp100[b]
Ureaplasma urealyticum	Bacterium	0.75	613	1(Wong & Houry, 2004)	1(Wong & Houry, 2004)	0(Wong & Houry, 2004)	0(Wong & Houry, 2004)	0	1(Wong & Houry, 2004)
Synechocystis sp. PCC 6803	Bacterium	4	3725	3(Dupre, Ruppr echt, & Schn eider, 2011; Ruppr echt, Duppr e, & Schn eider, 2010)	7	1(Fang & Barnum , 2003)	2(Lund, 2009)	1(Gie se & Vierli ng, 2002)	2
Mycobacteriu m tuberculosis	Bacterium	4.4	3987	1(Wong & Houry , 2004)	2(Wong & Houry, 2004)	1(Wong & Houry, 2004)	2(Lund, 2009)	2(Ken nawa y et al., 2005)	1(Wong & Houry, 2004)
Bacillus subtilis	Bacterium	4.2	4176	1(Wong & Houry , 2004)	1(Wong & Houry, 2004)	1(Wong & Houry, 2004)	1(Wong & Houry, 2004)	3(Rei schl, Thak e, Homu th, & Schu mann , 2001)	0(Wong & Houry, 2004)
Escherichia coli	Bacterium	4.6	4258	3(Vos , Hage man, Carra , & Kamp inga, 2008)	6(Vos et al., 2008)	1(Chen , Zhong, & Monteir o, 2006)	1(Lund, 2009)	2(Vos et al., 2008)	1
Bradyrhizobiu m japonicum USDA110	Bacterium	9.1	8374	1	4	1	5(Lund, 2009)	11(Le ntze, Aquili na,	2

								Lindbauer, Robinson, & Narberhaus, 2004)	
Methanocaldococcus jannaschii	Archaea	1.7	1715	0(Macario, Brocchieri, Shenoy, & Conway de Macario, 2006)	0(Macario et al., 2006)	0	1(Macario, Malz, & Conway de Macario, 2004)	1(Bult et al., 1996)	0(Wong & Houry, 2004)
Halobacterium sp. NRC-1	Archaea	2.6	2630	1(Macario et al., 2006)	2	0(Wong & Houry, 2004)	2(Macario et al., 2004)	3(Shukla, 2006)	0(Wong & Houry, 2004)
Sulfolobus solgataricus	Archaea	3.0	2977	0(Macario et al., 2006)	0(Macario et al., 2006)	0(Wong & Houry, 2004)	3(Macario et al., 2004)	2(D. C. Li, Yang, Lu, Chen, & Yang, 2012; Wang et al., 2010)	0(Wong & Houry, 2004)
Plasmodium falciparum	Protist	22.9	5428	6(Botha, Pesce, & Blatch, 2007)	43(Botha et al., 2007)	2(Chen et al., 2006)	11(Pavithra, Kumar, & Tatu, 2007) (9/2)	2(Pavithra et al., 2007)	1(Pavithra et al., 2007)
Dictyostelium discoideum	Protist	34	13212	9	38	2(Chen et al., 2006)	9 (8/1)	17	2
Cyanidioschyzon merolae	Algae	16.5	4998	4(Renner & Waters, 2007)	26	2	13 (10/3)	2(Waters & Rioflorido, 2007)	4
Chlamydomonas reinhardtii	Algae	121	14416	5(Renner & Waters, 2007)	63(Willmund, Dorn, Schulz-Raffelt, & Schroda, 2008)	3(Chen et al., 2006)	12 (8/4(Nordhues, Miller, Muhlhaus, & Schroda, 2010))	5(Waters & Rioflorido, 2007)	3(Nordhues et al., 2010)
Saccharomyces cerevisiae	Fungus	13	6692	16(Vos et al.,	22(Qiu, Shao, Miao, &	2(Chen et al., 2006)	9 (8(Fares & Wolfe, 2003)/1)	7(Gong et al.,	2(Gong et al., 2009)

				2008)	Wang, 2006)			2009)	
Neurospora crassa	Fungus	38	9820	12(Borkovich et al., 2004)	18(Borkovich et al., 2004)	1(Chen et al., 2006)	9 (8/1(Borkovich et al., 2004))	3(Borkovich et al., 2004)	2(Borkovich et al., 2004)
Caenorhabditis elegans	Nematode	103	20389	13	28(Zhao, Braun, & Braun, 2008)	3(Chen et al., 2006)	9 (8(Fares & Wolfe, 2003)/1)	18(Avermann & Waters, 2008)	0
Apis mellifera	Insect	236	10694	6	28	4(Chen et al., 2006)	9	10(Z. W. Li et al., 2009)	0
Drosophila melanogaster	Insect	180	13197	14(Vos et al., 2008)	36(Vos et al., 2008)	3(Chen et al., 2006)	12 (8(Monzo, Dowd, Minden, & Sisson, 2010)/4(Sarkar & Lakhotia, 2005))	11(Z. W. Li et al., 2009)	0
Arabidopsis thaliana	Land plant	120	27416	18(Vos et al., 2008)	116(Rajan & D'Silva, 2009)	7(Chen et al., 2006)	18(Hill & Hemmingsen, 2001) (9/9)	19(Waters, Aevermann, & Sanders-Reed, 2008)	3(Lee et al., 2006)
Populus trichocarpa	Land plant	485	41377	26	155	10	27 (17/10)	36(Waters et al., 2008)	4(Singh, Singh, Mittal, & Grover, 2010)
Oryza sativa	Land plant	389	57939	26(Hu, Hu, & Han, 2009)	125	7(Chen et al., 2006)	23	23(Waters et al., 2008)	3(Singh et al., 2010)
Ciona intestinalis	Ascidian	160	14180	8(Wada, Hamada, & Satoh, 2006)	36(Wada et al., 2006)	3(Chen et al., 2006)	11 (9/2)	6(Franck et al., 2004)	0
Gallus gallus	Vertebrate	1050	16736	10	44	4(Chen et al., 2006)	9 (8/1)	10(Franck et al., 2004)	0
Homo sapiens	Vertebrate	2851	20876	17(Brocchieri, Conway de Maca	41(Qiu et al., 2006)	5(Chen et al., 2006)	12(Mukherjee, Conway de Macario, Macario, & Brocchieri, 2010) (11/1)	11(Vos et al., 2008)	1

				rio, & Macario, 2008)					
Danio rerio	Vertebrate	1500	26100	10	68	6(Chen et al., 2006)	9 (8/1)	13(Elicker & Hutson, 2007)	1

[a]Genome size (in megabases) and the number of genes in each genome were taken from the relevant Enzembl database with the following exceptions: genome information for *Ureaplasma urealyticum* and the three Archaea was obtained from the UCSC Microbial Genome Browser; genome information for Synechocystis PCC 6803 was obtained from CyanoBase; and genome information for Bradyrhizobium japonicum USDA110 was obtained from RhizoBase.

[b]Numbers of chaperone genes were obtained from the indicated references. Where there is no reference, the numbers of chaperone genes were obtained from the relevant genome by identifying all of the genes that are annotated as having the correct type of fold. The numbers of Hsp60 subunits are divided between cytoplasmic and organellar forms. References were sometimes available for one form and not the other, as indicated.