COMPREHENSIVE ANALYSIS OF INTRINSICALLY DISORDERED PROTEIN CONTENT IN ORGANISMS EXPOSED TO EXTREME AMBIENT CONDITIONS

Maria Esmeralda Vicedo Jover

FAKULTÄT FÜR INFORMATIK

DER TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

FAKULTÄT FÜR INFORMATIK
DER TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Bioinformatik

COMPREHENSIVE ANALYSIS OF INTRINSICALLY DISORDERED PROTEIN CONTENT IN ORGANISMS EXPOSED TO EXTREME AMBIENT CONDITIONS

Maria Esmeralda Vicedo Jover

Vollständige Abdruck der von der Fakultät für Informatik der Technische Universität München zur Erlangung des akademisches Grades eines

Doktors der Naturwissenschaften

genehmigte Dissertation.

Vorsitzende:	Univ.-Prof. Gudrun J. Klinker
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. Burkhard Rost
	2. Prof. Dr. Bertram Müller-Myhsok University of Liverpool/UK

Die Dissertation wurde am 15.12.2015 bei der Technische Universität München eingereicht und durch die Fakultät für Informatik am 16.06.2016 angenommen.

Science is at times known
And at times unknown.
Nature is always
Unknowable.

(SCNT 14)

Science is revolution.
Nature is evolution.

(SCNT 15)

Science and nature
by Sri Chinmoy
Agni Press, 1996

Contents

Abstract **8**

List of publications and conferences **10**

1 Introduction **12**
1.1 Several IDP predictors for several flavors of disorder 12
1.2 Disordered regions associated with the complexity of the organism 13
1.3 Disordered regions and extreme organisms 14
1.4 Chromosomal duplication under stress situations 14
1.5 Objective and outline of the thesis 15

2 Materials and Methods **18**
2.1 Data 18
2.2 Quality measures of disorder predictions 20
2.3 Using Z-score to normalize disorder predictions 22
2.4 Homology based comparisons 23
2.5 Statistical methods 24
2.6 GO term enrichment 26
2.7 PPI interactions 27

3 Results and Discussion **28**
3.1 Protein disorder content differences in prokaryotes 28
3.1.1 Salt selects for high disorder 28
3.1.2 Heat selects for low disorder 28
3.1.3 Cold selects for high disorder 30
3.1.4 High disorder protects from radiation 31
3.1.5 Other outliers predictions for disorder content 31
3.1.6 Homology and association rules confirm the results 32
3.1.7 Statistical analysis of disorder content distribution: habitat vs. phyla 35
3.1.8 Null hypothesis rejected: disorder is similar between habitats 35
3.2 Protein disorder content under high temperature pressure 38
3.2.1 Duplications reduce the overall amount of protein disorder 38
3.2.2 Heat-shock proteins 41
3.2.3 GO terms enriched for growth and reproduction in heat stress-duplicate regions 43
3.2.4 Protein localization and disorder content 46
3.2.5 Protein-Protein Interactions (PPI) 47

4 Conclusion **50**

Bibliography **52**

Acknowledgements **66**

Appendix **68**

Abstract

Proteins that do not adopt well-defined tree-Dimensional (3D) structures in isolation are called intrinsically disordered proteins (IDPs). IDPs present some unique biophysical characteristics that allow them to bind to a wide range of partners. Binding to different partners may take place at different times and under different cellular conditions making their study even more difficult.

In one publication, we analyzed the content of IDPs in prokaryotes adapted to extreme habitats. Some differences in organisms surviving in extreme habitats correlate with a simple single feature, namely the fraction of proteins predicted to have long disordered regions. Moreover, the genomes of such extremophiles differ from their non-extremophile relatives. We began with the prediction of disorder with different methods for 40 completely sequenced prokaryotes from diverse habitats and found a correlation between protein disorder and the extremity of the environment. In particular, the overall percentage of proteins with long disordered regions tended to be more similar between organisms of similar habitats than between organisms of similar taxonomy.

Motivated by the above results, we analyzed the protein disorder content in a culture of *Saccharomyces cerevisiae* (baker's yeast) that survives sudden experimentally induced high temperatures by specifically duplicating the entire chromosome III and two chromosomal fragments (from chromosomes IV and XII). First, we established that heat shock proteins (HSPs) are not significantly over-abundant in the duplication, i.e. what many might have considered the simplest explanation failed. In contrast, a simple algorithm explains some of the experimental results: find a small enough chromosome with minimal protein disorder and duplicate this region. Additional analysis of functional involvement and networks of protein interactions added arguments for the observed duplication. Ultimately, it seems that the reduction of proteins with long regions of disorder allows *Saccharomyces cerevisiae* to decrease the effect of a heat shock attack.

List of publications and conferences

The work constitutes a cumulative dissertation. The methodologies and results as presented here - in particular sections 2.1-2.3 and 3.1-3.2 have been published in the following peer-reviewed articles. The manuscripts have been appended to this dissertation.

- **Esmeralda Vicedo**, Avner Schlesinger & Burkhard Rost, ***Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes***. PLoS One 2015, 10:e0133990.[1]

- **Esmeralda Vicedo**, Zofia Gasik, Yu-An Dong, Tatyana Goldberg & Burkhard Rost, ***Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock***. F1000Research 2015, 4:1222 (doi: 10.12688/f1000research.7178.1)[2]

The contents of the following peer-reviewed publications and conference attendances have not been used directly for the work.

- László Kaján, Guy Yachdav, **Esmeralda Vicedo**, […] and Burkhard Rost, ***Cloud Prediction of Protein Structure and Function with PredictProtein for Debian***, Hindawi Publishing Corporation, BioMed Research International, Volume 2013, Article ID 398968

- Tobias Hamp, Rebecca Kassner, Stefan Seemayer, **Esmeralda Vicedo**, [...] and Burkhard Rost, ***Homology-based inference sets the bar high for protein prediction***, BMC bioinformatics 2013, 14 Suppl 3:S7

- Predrag Radivojac, Wyatt T Clark, […],**Esmeralda Vicedo**, […],Christine Orengo, Burkhard Rost, Sean D Mooney, Iddo Friedberg, ***A large-scale evaluation of computational protein function prediction***, Nat Methods. 2013 Mar,10(3):221-7

- Avner Schlessinger, Christoph Schaefer, **Esmeralda Vicedo**, Markus Schmidberger , Marcos Punta, Burkhard Rost, ***Protein disorder breakthrough invention of evolution?***, Current Opinion in Structural Biology, 2011 Jun,21(3):412-8.[3]

- **Esmeralda Vicedo**, Avner Schlessinger, Burkhard Rost. ***Environmental pressure imprinted upon genomes through protein disorder***, 21th Annual International Conference on Intelligent Systems for Molecular Biology and 12th European Conference on Computational Biology 2013 (ISMB/ECCB), Berlin, Germany, July 21-23 2013 (Poster).

- **Esmeralda Vicedo**, Avner Schlessinger, Markus Schmidberger, Burkhard Rost. ***Are protein disordered regions equal to loops?*** 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology (ISMB/ECCB), Vienna, Austria, July 15-17 2011 (Poster).

1 Introduction

Most proteins adopt well-defined three-dimensional (3D) structures [4-7], *i.e.* predominately appear identical at different times. Opposed are intrinsically disordered proteins (IDPs) or regions (IDRs), which do not adopt well-defined 3D structures in isolation [3], *i.e.* they adopt several forms at different times when they are not binding to substrates. IDPs incorporate some exclusive biophysical properties allowing them to interact to an immense spectrum of partners, a number of times under distinct cellular conditions [8]. Proteins with IDRs seem to be exceptionally rich in processes such as transcription, translation, signal transduction, and macromolecular transport through the nuclear pore complex [9-11]. All these findings support the idea of disordered regions as elementary units for elaborate organizations [3].

1.1 Several IDP predictors for several flavors of disorder

Disordered regions are not so easy to observe experimentally. There is not one ideal experimental method able to capture all proteins in an organism with long regions of disorder (*e.g.* regions not observable in X-ray crystallography, signal overlap [12]). Instead, acceptably precise computational predictions are available for whole proteomes [13-15]. Experimental methods are frequently restricted by low spatial resolution and labeling complications [16-18]. Another problematic point is the confusing definition of IPDs, as it covers every protein that does not presents a well-defined three-dimensional structure. With the result that the existing prediction methods were designed around diverse concepts of disordered proteins and therefore each of them captures several types or "flavors" of protein disorder (*e.g.* extended, collapsed, or combinations of both) [19-21]. Thus, when examining the number of predicted disordered proteins in an organism, this might be different depending on the predictor.

In our analysis therefore we applied three completely different predictors (namely IUPred, Meta Disorder and NORSnet), to overcome the potential biases of the predictors. IUPred uses pairwise statistical potentials of residue contacts [22, 23] and has been described as an unbiased and robust predictor even for organisms living in extreme conditions [24, 25]. Meta-Disorder (MD) [21] and NORSnet [20] are neural network-based methods which include evolutionary information and other predicted features in their predictions. More precisely, MD is a predictor which combines apart from several original prediction methods (including NORSnet), evolutionary profiles and sequence features correlated with protein disorder such as predicted solvent accessibility and protein flexibility (22). The other appointed predictor method,

NORSnet, is more focused on the identification of long disordered loops (no regular secondary structure, specifically "loopy disorder") (10). Furthermore, NORSnet was optimized without using any experimental disordered protein data (10) which ensures a certain degree of independence with respect to other methods based on experimental data. Generally, disordered regions that are not predicted to be "loopy" are considered "regular" disordered regions. Using NORSnet it is possible to discriminate "loopy" disorder from the rest of several disordered protein groups [19].

1.2 Disordered regions associated with organism's complexity

At the category of kingdoms, 10-20% of all proteins from prokaryotes have at least one long disordered region, while for eukaryotes it is at least 20-50% of all proteins [3, 26, 27]. Latest comparative proteomics studies have intensified the association between protein disorder and organism complexity, *e.g.* disordered regions in "younger" ramifications of eukaryotes seem to be different from "older" braches of eukaryotes [28-30].

One way to compare genomes is by analyzing characteristics of proteins. For example, combining the analysis of sequence, structure, expression and evolutionary relationship information of multiple protein data sets from several organisms (*e.g.* yeast, mouse and human) allows to find evidence about the possible connections between disparity in the length of disordered regions and modifications in the protein functions among the organisms [31]. Changes in the length of disordered regions in paralog proteins might supply a simple evolutionary instrument for measuring protein degradation rates. Furthermore, frequently many of these affected paralogs were participating in protein signaling pathways which also influence the cellular function and phenotype of the cells [32-34]. On ordered proteins, the secondary structure elements (α-helix, β-strand and coil) can be easily visualized as protein crystal structures and predicted from the amino acid composition [35, 36] and homology domain [37, 38]. It is also recognized that intertwined helices (coiled-coils) are abundantly present in eukaryotes [39] and that they are built with local internal molecular interactions [40]. For all of that, helices are considered as evolutionary building blocks and this useful information can be integrate to implement prediction methods to study structural characteristics of entire proteomes within species [10, 37, 39, 41-44].

1.3 Disordered regions and extreme organisms

Protein disorder seems to be one means for prokaryotes to adapt to extreme environments, *e.g.* thermophiles have much fewer proteins with long disorder than their closest phylogenetic relatives living in mesophilic conditions. It appears intuitive to assume that increasing the internal inter-residue bonds in a protein raises its stability at high temperature. Diverse studies have, effectively, reported association between thermal stability and "order related" attributes in proteins such as a high contact density and hydrogen bonds [45, 46]. Moreover, when considering in more detail the amino acid composition of proteins from thermophiles and mesophiles a difference in the average amino acid composition was found [47]. Protein structures from thermophiles such as *Pyrococcus horikoshii OT3* have been reported to contain more intra-helical salt bridges than their homologues in mesophiles [48]. These salt bridges are an important factor stabilizing thermophilic proteins [46].

All these findings suggest that diverse factors determine thermostability [49]. On the other hand recent studies of psychrophiles, organisms living at the opposite end of the temperature scale of habitats, namely in the extreme cold, have suggested that proteins from psychrophiles increase their flexibility and accessibility and thereby might hinder freezing [50]. Moreover, proteins from Halobacteria (salty habitats) also exhibit unique characteristics such as low hydrophobicity, excess of acidic residues, depletion of cysteine residues and reduced propensities for helix formation [51]. In the light of this information we would ask: Is there any evidence to suggest that disordered regions in proteins may increase the fitness of extremophiles?

1.4 Chromosomal duplication under stress situations

Saccharomyces cerevisiae (baker's yeast; we mostly used the abbreviation yeast) was the first completely sequenced eukaryote [52] due to its widespread use in the experimental studies as model organism [53-55]. Presents a restricted temperature range for optimal growth but tolerates moderate deviations, some of which affect cell structure and function, often through rapid physiological adaptations. One of such adaptation instrument is the duplication of the whole genome or specific chromosomes (aneuploidy) [56-58] that contain the genes necessary to quickly respond to the particular unfavorable circumstances through many generations of evolving yeast [59-65]. Such reactions to the environment imbalance the genome [66], destabilize some reactions and pathways [67, 68] and appear to cost substantial energy [69, 70]. Aneuploidy, therefore, is a temporary response that is replaced by specific refined and less expensive solutions, when yeast is exposed to the same adverse environment over many generations [71]. A recent experiment established that yeast

cells evolving under high temperature can adapt to this stress by duplicating chromosome III as well as fragments of chromosomes IV and XII [71]. However, the underlying reason for this response to high temperature, *i.e.* specifically copying these regions, remains as yet unknown.

1.5 Objective and outline of the thesis

The primary objective of this thesis was to analyze the protein disorder content of several organisms exposed to extreme environments in order to find possible correlations between protein disorder and the extremity of the environment. For that we combined all experimental data available at the moment of this work (complete proteome, genome and metadata [72]) and several methods, tools and additional prediction data (*i.e.*, homology, GO enrichment, protein-protein interaction, subcellular localization prediction data) . Firstly, I presented in the thesis a more generalized overview as we analyzed the content of intrinsically disordered proteins in several organisms living in different environments (. Then, I went into more detail and analyzed the disorder protein content and duplication of chromosomes (aneuploidy) in a model organism (yeast) exposed to a sudden extreme condition (high temperature stress).

In the first part of the thesis (sections 2.1, 2.2 and 3.1), we combined the protein disorder predictions with the experimental environmental information (optimal temperature, pH, habitat, among other important aspects) and taxonomical information. We picked out several organisms classified as extreme (living in extreme conditions) and their taxonomical neighbors living in non-extreme environments in order to compare their protein disorder content. As the available experimental information was not enough to explain the observed differences in protein disorder content, we also introduced two different approaches, namely homology (section 2.4) and statistical tests comparisons (section 2.5).

In the second part (section 3.2), we aimed to explain the aneuploidy (chromosomal duplication) in *Saccharomyces cerevisiae* (yeast) under induced high temperature stress from new computational perspectives, specifically the protein disorder content. As that was not enough to explain these phenomena, we then combined all experimental information available about yeast with computational methods, predictors and tools (GO enrichment, protein-protein Interaction, subcellular localization). This allowed us to give possible answers and show several possibilities for further experimental research (sections 3.2.1-3.2.5).

In both cases, we used different predictors and methods (section 2.2) to avoid introducing bias by the predictors themselves due to the extreme conditions of the environments to which the proteins were exposed. Most importantly, the predictors

were not designed nor tested to predict proteins of organisms affected by extreme conditions.

2 Materials and Methods

2.1 Data

The UniProt database [73] delivered the complete proteome sequence data for our study. We discharged all duplicated protein (giving priority to longer sequences). The first analysis considered organisms living in extreme conditions and their closes relatives with a total of 46 organisms and 225,550 proteins[1] (Table 2.1). The organisms sampled the most extreme habitats and any of their closest completely sequenced relatives. In total 19 organisms living in extreme environments are considered. In addition 21 mesophile organisms living in "normal" environments that are related to the 19 extremophiles are analyzed. We also included a few selected eukaryotes living in normal environment for comparison.

Most information used to classify organisms was taken from GOLD (Genomes OnLine Database, version 2011-09-23 [72]). We avoided pathogens, parasites, and other biotic relationships to build a "simplified" subset of organisms. We classified the organism into the following types of environment (Table 2.1) [74-76]: thermophiles (optimal growth at 45-80º C), hyperthermophiles (temperature optima >80º C), psychrophiles (optimal growth at about 15º C, a maximal temperature for growth at about 20º C, and a minimal temperature for growth at 0º C or below), psychrotolerants (organisms that are not considered as psychrophiles but have the capability for growth at 0º C or close to 0º C), halophiles (optimal growth in salt solutions, *i.e.* from 25% NaCl up to saturation), alkaliphiles (optimal growth around pH>8), mesophiles (including bacteria and archaea from "normal" environments). Eukaryotes were considered as a different group as they have a different content of disorder [3].

For the second part of the analysis, the complete proteome of yeast was downloaded from the UniProt database (www.uniprot.org, release 2013-10.) [73]. The duplicate proteins were removed (considering 100% pairwise sequence identity and keeping the longer sequence) which left just 5667 proteins [2]. Only the 16 autosomal nuclear chromosomes were considered (matched through the Saccharomyces Genome Database abbreviated as SGD, www.yeastgenome.org [77]; Fig. 2.1), while the allosomes (also referred to as sex chromosomes) and mitochondrial DNA were not included in the study. The SGD database also provided the annotations of heat-shock response (HSR) proteins. All proteins known to interact with HSR proteins were added to this set of HSR proteins.

Table 2.1: List of organisms grouped after environmental conditions [1]. Listed are for each environmental condition (here Organism groups), the number of organisms, the percentage of organisms, the number of proteins for this environmental condition/in this group and the percentage of proteins. In total 19 organisms living in extreme environments are considered. In addition 21 mesophile organisms living in "normal" environments that are related to the 19 extremophiles and 6 eukaryotes are analyzed.

Organism groups	# Organisms	% Organisms	# Proteins	% Proteins
Thermophiles	3	6.52	7781	3.35
Hyperthermophiles	2	4.35	3746	1.66
Psychrophiles	3	10.87	12918	5.73
Psychrotolerants	4	4.35	14044	6.23
Halophiles	3	6.52	10772	4.78
Alkaliphiles	1	2.17	3981	1.77
Radiation resistant	3	6.52	9759	4.33
Total extreme organisms	19	41.3	63001	27.93
Mesophiles (Bacteria + Achaea)	21	45.65	70954	31.46
Eukaryotes	6	13.04	91595	40.61
Total organisms	46	100.00	225550	100.00

The data for experimental protein-protein interactions (PPIs) in yeast was provided by BioGRID (version March 2012). After filtering out repetition (a-b and b-a counted only once) and self-interactions (a-a), we based all subsequent analyses on the single largest connected component of the network.

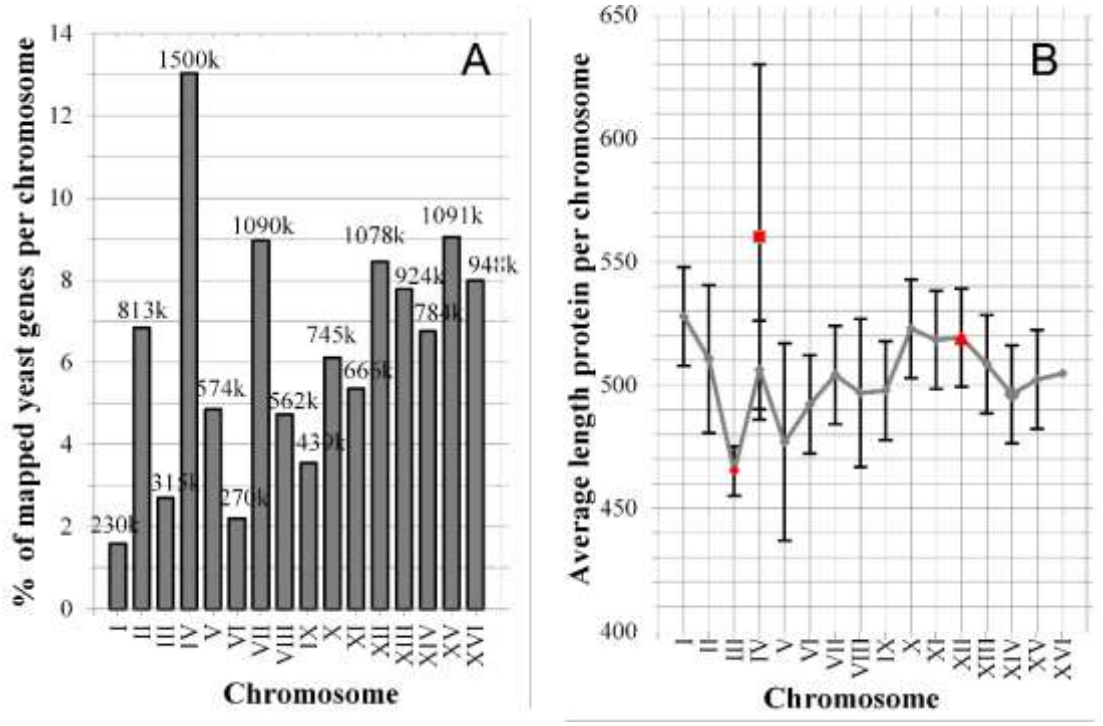

Figure 2.1: Number of genes per chromosome in *Saccharomyces cerevisiae* [2]. (A) Distribution of the percentage of genes mapped to each of the 16 chromosomes of yeast (*Saccharomyces cerevisiae*). The numbers on top of the bars mark the length of the chromosome in kilobase pairs (marked as k in the graph). (B) Distribution of average protein length in each chromosome. The red dots mark the chromosomes III, IV and XII that are affected by duplication (whole chromosome duplication in chromosome III and duplication of fragments in chromosomes IV and XII).

2.2 Quality measures of disorder predictions

We applied distinct prediction methods in order to capture several "flavors" of protein disorder [19-21], obtaining distinct values of disorder content depending on the predictor. IUPred utilizes pairwise statistical potentials of residue contacts [22, 23] and has been reported as an unbiased and robust predictor even for organisms living in extreme habitats [24, 25]; Meta-Disorder (MD) [21] and NORSnet [20] are neural network-based methods that utilize evolutionary information as well as other predicted features. MD incorporates several original prediction methods including NORSnet into evolutionary profiles and sequence features correlated with protein disorder (such as predicted solvent accessibility and protein flexibility among others aspects). NORSnet is based on long disordered loop identification (no regular secondary structure, namely "loopy disorder"). One advantage of this method is its optimization without using any experimental data based on disorder. We considered

the disordered regions that are not predicted to be “loopy” as “regular” disordered regions.

There are many approaches how to compute overall protein disorder content in proteins (*e.g.* content of IDR, % amino acids in the sequence considered as disorder [15]. We analyzed almost the entire resulting flood of data and found most of the alternatives to be superfluous. Therefore, we decided to focus our studies on few alternatives; we included these different views only if they provided important additional information. We also introduce a radical concept to classify a protein as disordered, in particular the approach to define completely disordered proteins.

Long disorder. We generally defined “long disorder” using one threshold: **%long30**, *i.e.* the percentage of proteins with at least one region of ≥30 consecutive residues predicted as disordered (other alternatives: **%long80** and **%long50** with length thresholds of ≥80 and ≥50 consecutive residues, respectively).

Completely disordered. If a protein had no single region that we could identify as a regular structure, we considered this protein as completely disordered (Fig. 2.2). We reported the fraction of all proteins that fit this criterion. Basically, the procedure consist of these steps: first, remove predictions of disorder that cover fewer than five residues; next, search regions of ≥30 residues or more with all residues predicted as not disordered. If no such region is found, but the opposite, namely at least one region with ≥30 consecutive residues predicted as disordered, we consider the protein to be completely disordered (Fig. 2.2). All thresholds (region length ≥30, ≥50, ≥80) were tested using the three prediction methods (MD, NORSnet and IUPred).

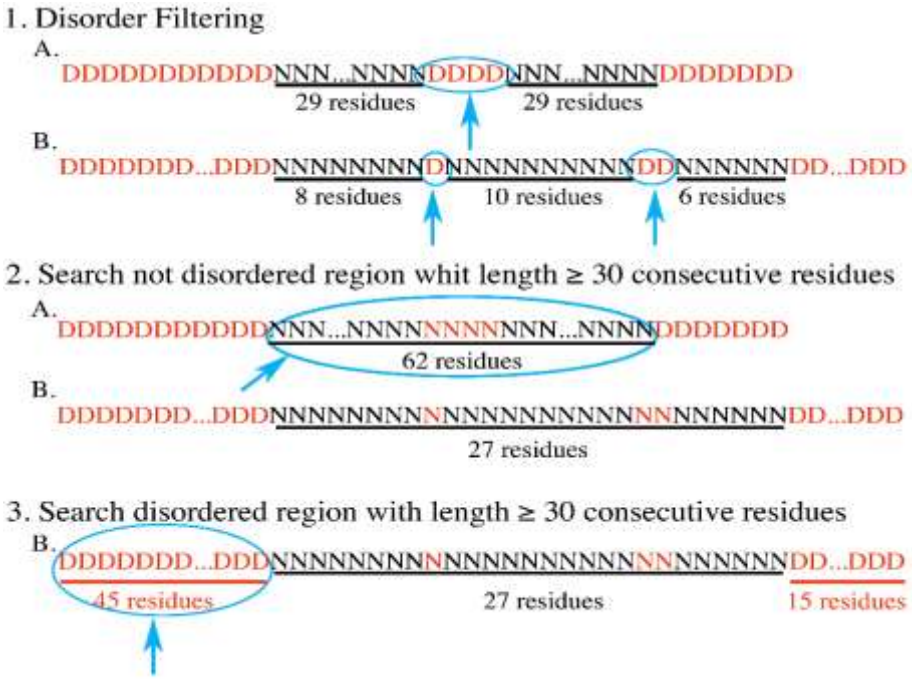

Figure 2.2: Processing steps for "completely disordered" approach [1]. This method was devised to capture proteins without a single region that we could perceive as a "nucleation site" for adopting regular structure. Operationally, we first removed any prediction of disorder that spanned over fewer than five residues (step 1); next we searched any region without predicted disorder over 29 consecutive residues (step 2). If we found no such region, but we found at least one region with ≥30 consecutive residues predicted as disordered, we considered the protein as "completely disordered" (step 3).

2.3 Using Z-score to normalize disorder predictions

To facilitate the correlation between the outputs of the three predictor methods when studying complete proteomes, we substituted their raw scores by Z-scores, *i.e.* we gave the score as a deviation from the average in units of one standard deviation [1]:

$$z(o,M) = \frac{raw(o,M) - \langle raw \rangle (\text{all organisms}, M)}{\sigma(\text{all organisms}, M)} \qquad \text{(Eqn. 1)}$$

where z(o, M) is the Z-score for a particular method M and organism o, raw(o, M) is the raw score of the prediction method M for organism o (*e.g.* the percentage of proteins with at least one region of long disorder in o), <raw>(allorganisms, M) is the average of the raw scores for method M over all organisms, and σ(allorganisms, M) is the standard deviation for the distribution of the raw scores predicted for all

organisms by method M. Positive Z-scores signified a disorder content higher than the mean, while negative scores signify disorder content lower than the mean.

We collected averages and standard deviations over a set of 1613 complete prokaryotic proteomes taken from UniProt (release 2013-10; we only considered the organisms with almost 90% (≥ 90%) of the sequences predicted by the three predictors [1]. We used this score in order to have a Z-score calculated independently of the samples selected and to present the information over the total amount of complete proteomes available [1] (Fig.2.3). Eukaryotes were excluded in this computation of Z-score as their disorder content [13] differs substantially from that of prokaryotes. They were considered as an independent group.

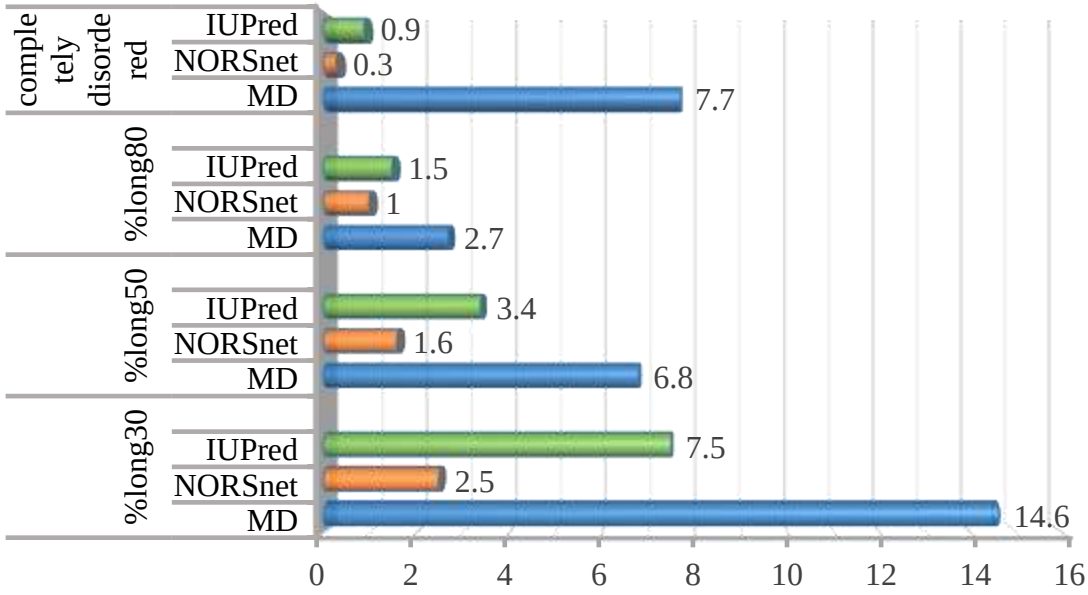

Figure 2.3 Protein disorder content for prokaryotes [1]. We computed the means and standard deviations to calculate the Z-score for the three predictors (MD, NORSnet and IUPred) and four methods (%long30, %long50, %long80 and completely disordered) using the total of complete proteomes available in our data (1613 prokaryotes).

2.4 Homology based comparisons

To identify phylogenetic relations between two organisms, we applied the following *ad hoc* procedure: we ran all protein sequences from one organism against all from the other using PSI-BLAST [78]. Afterwards, for each alignment obtained we calculated the HSSP-value (HVAL) [38, 79, 80], which quantifies sequence similarity by combining alignment length and percentage of pairwise sequence identity. For

example, HVAL=0 is equal to about 22% pairwise sequence identity for alignments over 250 residues. Our concept allowed proteins to have multiple homologues. Due to technical difficulties, we did not discriminate between paralogs and orthologs [81, 82].

2.5 Statistical methods

To collect the possible differences in the disorder content between organisms with a similar habitat (Table 2.1) as well as those of similar phylogeny (Table 2.2), we applied several statistical tests, namely the Kruskal-Wallis test (also called as H-test) [83, 84], the Wilcoxon signed-rank test [85-88] and a robust Brown–Forsythe Leven's test of equality of variances (shortly named Leven's test). The Kruskal-Wallis test is a non-parametric overall statistical method used to measure the shape of the distributions of two or more unmatched groups. It is applied on nominal variables of small and unequal sample size to determine whether the distributions of the groups (protein disorder content distribution) are identical (null hypothesis) [83, 85, 86]. The Kruskal-Wallis test does not assume a normal distribution for the data but homoscedasticity (not significant differences between the group variances). The Wilcoxon signed-rank test (called Wilcoxon test) is a pairwise statistical test applied to assess whether the means of two groups differ. The Leven's test is a non-parametric test based on the medians of the groups [89, 90]. It is a test that is robust against deviations from normality and expects equality for all group variances (null hypothesis).

First of all, we performed a Leven's test, to avoid misleading the homoscedasticity premise of the Kruskal-Wallis test (Fig. 2.4). If the Levene's test succeeded for the overall comparison across the groups, *i.e.* the null hypothesis is accepted (equal variances), then we applied the Kruskal-Wallis test (groups have equal distribution). In case, the overall Levene's test failed, we performed pairwise comparisons between the groups (pairwise Levene's test). For those groups for which the null hypothesis was accepted, then as an alternative to the Kruskal-Wallis test, we applied a pairwise Wilcoxon test (Fig. 2.4).

Moreover, we also applied the pairwise Wilcoxon signed-rank test when the Kruskal-Wallis test failed for the overall comparison test (acceptance of the alternative hypothesis, *i.e.* at least one group in the population for which the distribution of disordered protein contents differs from the others; Fig 2.4).

Table 2.2: List of organisms grouped after similar phylogeny [1]. Listed are the organisms grouped after the taxonomy classification used by NCBI database (here Phylogeny), the number of organisms involved in each group (here Organisms) wherein groups marked with an asterisk are considered for the statistical computation. It is also listed the number of extreme organisms in a particular group (here Extreme organisms) wherein the names and amount of extreme organisms in each group are between brackets Abbreviations used: **alkalo**, alkaliphiles; **thermos**, thermophiles; **hyperthermo**, hyperthermophiles; **pyschrotol,** psychrotolerants.

Phylogeny	**Organisms**	**Extreme organisms**
Actinobacteridae	2*	0
Alphaproteobacteria	5*	0
Bacilli	6*	3 (alkalo, pyschrotol, thermo)
Betaproteobacteria	3*	1 (pyschrotol)
Chroococcales	3*	1 (thermo)
Clostridia	2*	1 (thermo)
Deinococci	3*	3 (radio res)
Deltaproteobacteria	4*	1 (psychro)
Gammaproteobacteria	5*	4 (psychro, halo, psychrotol, psychro)
Halobacteria	2*	2 (halo)
Methanococci	1	0
Methanomicrobia	2*	1 (psychrotol)
Thermococci	1	1 (hyperthermo)
Thermoprotei	1	1 (hyperthermo)
Total	40 (37*)	19 (17)

The habitat is a complex world defined by a variety of ambient conditions and organism properties which have to be considered alone. That is why we also analyzed some general properties of the organisms (called metadata) which are included from the GOLD database [72]. Groups containing less than two samples were not included in the statistical analysis. We used the R software (statistical packages car and stats) [89, 91] to conduct all our statistical tests.

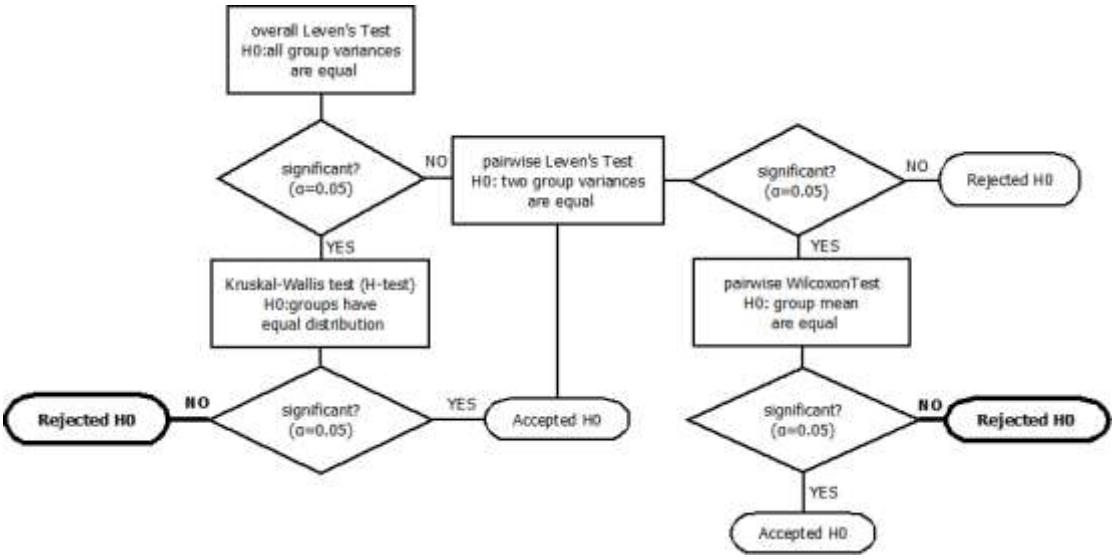

Figure 2.4: Graphical representation of the statistical analysis [1]**.** First of all, we performed a Leven's test, to avoid misleading the homoscedasticity premise of the Kruskal-Wallis test. If the Levene's test succeeded for the overall comparison across the groups, then we applied the Kruskal-Wallis test (groups have equal distribution). In case, the overall Levene's test failed, we performed pairwise comparisons between the groups (pairwise Levene's test). For those groups for which the null hypothesis was accepted, then as an alternative to the Kruskal-Wallis test, we applied a pairwise Wilcoxon test. We finalized the statistical test, when the null hypothesis (H0) of the Kruskal-Wallis Test or the pairwise Wilcoxon Test was rejected.

2.6 GO term enrichment

The Gene Ontology (GO) [92] represents protein function in three alternative branches. These are: biological process, molecular function, and cellular component. We focused our study on biological process and molecular function using the BINGO [93] application. We used this tool in our analysis to assess if the observation of a particular set of biological functions (to be more exact: GO numbers) was more statically significant than any other set. We used BINGO a Cytoscape plugin [94] to visualized the results. BINGO offers three tests for statistical significance: hypergeometric, Fisher, and binomial. Our analysis considered the hypergeometric test (a test without replacement). Using this test, we obtain more accurate p-values than using the other methods. We followed the common procedure [94] of considering p-values >0.05. We have to consider that testing multiple hypotheses may produce many false positives (Type I error: incorrect rejection of true null hypothesis [95, 96]). To corrected these false positive, we used the Benjamini and Hochberg correction

which allows strong control over the False Discovery Rate (FDR, expected proportion of erroneous null hypothesis rejections among all rejections [96]).

2.7 PPI interactions

Network biology provides a measurable description for networks describing multiple biological systems. We studied the most elemental network attributes that allow the contrast and definition of complex networks which are **degree**, **betweenness** and **average degree of neighbors**. **Degree**, also called connectivity, is the most elementary characteristic of a node and it is defined as the number of interactions for a given node. **Betweenness** is the fraction of shortest paths between all other nodes that has to go through a given node. **The average degree of neighbors** quantifies the number of nodes and links in the network. These three parameters measure the importance of each node within the network. In our case, a node was equivalent to protein.

3 Results and Discussion

3.1 Protein disorder content differences in prokaryotes

Here, we demonstrate that some differences in organisms surviving in extreme habitats (*i.e.* salt saturated environments, extreme heat or cold conditions, high radiation) correlate with the fraction of proteins predicted to have long disordered regions. In particular, we observe that the overall percentage of proteins with long disordered regions tends to be more similar between organisms of similar habitats than between organisms of similar taxonomy.

3.1.1 Salt selects for high disorder

The organisms living in salt saturated environments are called halophiles. We studied two halophilic archaea: *Halobacterium sp.* NRC-1 [97] and *Haloarcula marismortui* ATCC 43049 [98], for which around 20-28% of all proteins were predicted with long disorder (≥30 consecutive residues predicted to be disordered) by MD and IUPred [1]. This value was more elevated than the average for all prokaryotes (Z-scores Fig. 3.1A) and much higher than the values for their nearest taxonomic relative that does not survive in high salt: Methanococcus maripaludis S2 [99] (1-13% [1]). The same trend was noticed for the other methods [1].

The difference in disorder abundance between the halophilic bacterium *Marinobacter aquaeolei* VT8 [3] (Z-score around 0 [1],, Fig. 3.1A) and its taxonomic relative *Pseudoalteromonas atlantica* T6c [100] (Z-score around -0.5[1], Fig. 3.1A) was not as evident as for the archaea, but it also confirmed the "high disorder in salt habitat" trend for bacteria. The difference in disorder between halophile and relative was slightly higher for longer disorder [1]. The difference increased when considering the percentage of proteins considered as completely disordered [1]. If we consider the predictions of NORSnet, the method that detects only long loops (defined as no regular secondary structure) as disorder, the difference was the same in relative terms although the content for that method dropped significantly [1] (Fig. 3.1A). These observations across different phyla might suggest the increment in disordered regions as a way for prokaryotes to manage high salt conditions. This result has already been reported before [24, 101]. However, in our study we introduced the novelty of connecting phylogeny (closest relatives) to extremity of habitat (high salt) [1].

3.1.2 Heat selects for low disorder

Organisms surviving in extreme heat conditions present a lower disorder content ([24]). Our study was centered on two hyperthermophile archaea: *Pyrococcus* [102]

and *Aeropyrum pernix* K1[103-105]. *Pyrococcus* might be the most studied organism living in very high temperature (close to 100ºC). In addition to the high temperature, it is living in greater sea depth than other archaea (pressures reaching 200 bars, *i.e.* ~200 times what we live in). *Pyrococcus horikoshii* OT3 [106] presented nearly no long disorder (region of ≥30 residues, <-1, *i.e.* over one standard deviation below average [1], Fig. 3.1) at least by the predictions of two of the applied methods (MD and IUPred). When we observed the predictions for the closest relative, *Methanococcus maripaludis* S2, presented a disorder score similar to its hot relatives (Z-score around -1 [1], Fig. 3.1). The optimal growth temperature for *Methanococcus maripaludis* is 35-40ºC, *i.e.* "normal". Moreover, *Methanococcus* was isolated from salt marsh sediments. Following our simple logic, we propose two reasons for Methanococcus to have higher disorder than *Pyrococcus*: salt (higher disorder) and less heat (higher disorder). For our method predicting loopy disorder, the trend was instead inversed. We failed to explain why we did not observe that.

The other hyperthermophile, *Aeropyrum pernix* K1 was isolated from sulfur-rich undersea vents in Japan [103-105]. Like *Pyrococcus*, *Aeropyrum* was predicted with very low disorder content (Z-score ~-1 [1], Fig. 3.1), like two other hyperthermophiles that we sampled. Analogous to the halophiles, the "loopy" disorder predicted by NORSnet, was even lower for these hyperthermophiles than the "regular" disorder. We could suspect that shortening connections between regular secondary structure segments (helices and strands) might protect against heat and high salt but we should be careful with these speculations because this seems incompatible with the prediction of "loopy disorder" for *Pyrococcus* [1] (Fig. 3.1).

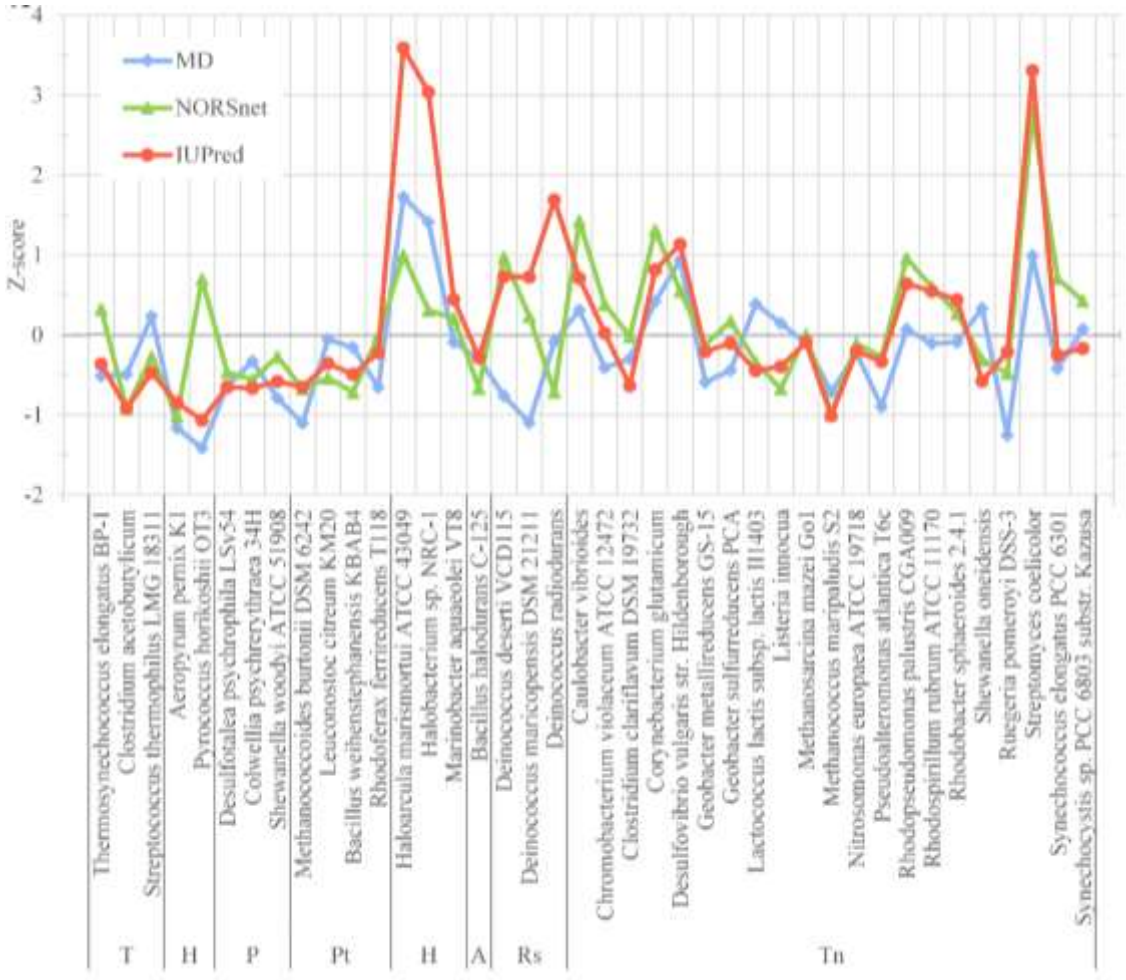

Figure 3.1: Distribution of disorder content in different organisms [1]**.** The graph represents the fractions of proteins with long regions (≥30 consecutive residues) of disorder predicted by three prediction methods (MD, NORSnet and IUPred). The raw values are standardized using the Z-scores (Eqn. 1). Eukaryotes were not included in this analysis. In the x-axis organisms are grouped after habitat where T denoted the thermophiles, H hyperthermophiles, P psychrophiles, Pt psychrotolerants, A alkaliphile organism, Rs radiation resistant organisms and Tn taxonomical neighbors of the listed extreme organisms.

3.1.3 Cold selects for high disorder

Colwellia psychrerythraea 34H [107], an obligate psychrophile marine bacterium, needs very low temperatures: -1ºC to +10ºC to grow. Moreover, it can support high pressures in the deep sea. Its predicted disorder was below the average (Z-score about -0.5 [1], Fig. 3.1). Another organism, *Leuconostoc citreum* KM20 [108] is considered to be a psychrotolerant antimicrobial producer (used for fermentation of kimchi). It grows optimally at 30ºC, but can also be cultivated at significantly higher

temperatures. Its predicted disorder was also below average (Z-score about -0.5 [1]; Fig. 3.1).

A new research presented experimental information that proteins with long disordered regions can be more stable in cold temperatures than globular proteins [109] which seemed incompatible with the concept that such a solution would be imprinted upon the entire proteome. But our analysis of psychrophiles [1] confirmed previous findings that organisms in cold habitats have less disorder than average [24]

3.1.4 High disorder protects from radiation

Deinococcus radiodurans R1 [110, 111] has been nicknamed as "Conan the bacterium" [112] because it tolerates many extreme conditions including radiation, cold, dehydration, heat and high acidity. For this bacterium (Z-score between 0 and 2[1]: Fig. 3.1), we predicted a high abundance of protein disorder. The two taxonomic neighbors of Deinococcus radiodurans, sustain also high radiation and live in the dry: Deinococcus deserti and Deinococcus maricopensis (Z-scores >0 for IUPred [1], Fig. 3.1). The 'high radiation' habitat was particularly inconsistent between the three prediction methods: e.g. MD predicted the opposite of the other two predictors [1] (Fig. 3.1). Inconsistency between prediction methods might suggest taking the correlation 'high radiation - high disorder' carefully. Contrary, we might claim the opposite: IUPred, MD, and NORSnet are based on independent and incomplete information which might imply that only the best of the methods might discover some reality. However, at present this is difficult to verify.

3.1.5 Other outliers predictions for disorder content

Finally, we analyzed the disorder content in prokaryotes that live in other extreme habitats including high pH (*Bacillus halodurans* [113], which presents a disorder content below average [1], Fig. 3.1A) and changing environments (*Shewanella oneidensis* [114], disorder around average [1], Fig. 3.1.). However, so far we failed to notice significant trends (Fig. 3.1). Moreover, we were unsuccessful in clarifying why some mesophiles were outliers (higher or lower content of disordered proteins). For example, *Caulobacter vibrioides* (also known as *Caulobacter crescentus*) [79] was predicted with high disorder (Z-score one standard deviation above average [1], Fig. 3.1) without any apparent reason. *Caulobacter* secretes Nature's strongest glue [115, 116] which might point to another important role for high content of disorder. *Streptomyces coelicolor* was also predicted with higher than average disorder (Z-score >1 [1], Fig. 3.1); that might be explained by its complex life cycle and production of antibiotics (its products are pharmaceutically used as anti-tumors agents, immunosuppressants and antibiotics).

Ruegeria pomeroyi DSS-3 [117] (originally classified as *Silicibacter pomeroyi* [118]) was predicted to have very low disorder (Z-score about -1 [1], Fig. 3.1). Its taxonomic neighbor, *Rhodobacter sphaeroides* 2.4.1, was predicted to have above

average disorder (Z-score>0 [1], Fig. 3.1). *Ruegeria* was isolated from seawater off the Southeast coast of the USA. It lives at 10-40ºC and grows with and without carbon monoxide (CO) as carbon source. We could not clarify the low protein disorder content predicted for *Ruegeria*.

3.1.6 Homology and association rules confirm the results

We calculated disorder abundance for only the homologue proteins of two model organisms which represent two extreme temperature environments (cold and hot) [1]. The homology was defined for various thresholds in terms of sequence similarity (Table 3.1). The purpose of this analysis was to examine the possible inclusion of disordered regions in the aligned region of the corresponding homologue proteins when considering two opposite extremophiles (heat/cold), specifically the *Colwellia psychrerythraea* 34H and the *Pyrococcus horikoshii* OT3. For instance, 24% of all 4423 *Colwellia* proteins have a match in one of the 1573 *Pyrococcus* proteins at HVAL≥0 (e.g. HVAL=0 implies 20% pairwise sequence identity (PIDE) for >250 aligned residues [80] (or 20+N% PIDE at HVAL=N), while almost 31% of the Pyrococcus proteins have a homolog in Colwellia at this level of sequence relation[1]. Overall MD predicts 12% of all *Colwellia* and 8% of all *Pyrococcus* proteins to have at least one long disordered region[1]. These numbers imply that the proteins shared between the two extremophiles from opposite ends of the temperature spectrum are depleted in disorder with respect to the entire proteome. For instance only 4.9% are related and disordered from the *Colwellia* perspective at HVAL≥0 as opposed to 12% for all proteins[1]. The more similar the homologs the more the related proteins were selected to not contain disorder.

Table 3.1: Protein disorder overlap between related proteins in opposite extremophiles[1]. The column "HVAL" measures sequence similarity as the distance from the HSSP-curve [38, 119]. The column "related" gives the percentage of proteins in one organism that have corresponding homologs in the other at the given HVAL (total of proteins: *Colwellia psychrerythraea* 34H=4423 and *Pyrococcus horikoshii* OT3=1573). The columns "related+disordered" gives the percentage of proteins in one that are "related" and have at least one disordered region (≥30 residues, prediction by MD; other methods and thresholds [1]) in the other organism at the given HVAL. One standard error is marked as '±stderr'.

HVAL	***Colwellia psychrerythraea* 34H (freeze)**		***Pyrococcus horikoshii* OT3 (heat)**	
	related	**related+disordered**	**related**	**related+disordered**
-20	75.5 ± 0.2	9.5 ± 0.1	66.9 ± 0.1	5.53 ± 0.06
-10	56.4 ± 0.2	6.8 ± 0.1	55.7 ± 0.2	5.04 ± 0.08
0	24.0 ± 0.1	4.9 ± 0.2	30.9 ± 0.1	2.7 ± 0.1
10	5.5 ± 0.1	2.6 ± 0.2	9.7 ± 0.1	0.51 ± 0.06
20	0.60 ± 0.02	0.07 ± 0.02	1.28 ± 0.03	0
30	0.04 ± 0.01	0	0.20 ± 0.01	0

We found, at pairwise protein similarity levels of HVAL≥10 (which correspond to about 30% pairwise sequence identity over 250 aligned residues), seven of the homologs with disorder in *Colwellia* (cold) have no disorder in *Pyrococcus* [1]. Instead, we detected only one protein with disorder in *Pyrococcus* and not in *Colwellia*.

Diverse studies analyzing the effect of temperature on enzymes – proteins which are disorder reduced - proposed that proteins from extremophiles (both cold and hot) assume similar structures as their mesophilic orthologs, but use distinct amino acids composition to accommodate for temperature effects [46, 47, 50]. Our analysis confirmed this trend [1] (Fig. 3.2A). However, the differences were significant at best for some particular amino acids. The strongest signal was for negatively charged amino acids such as glutamic acid (E, [1] Fig. 3.2A), that occurred more in heat than in cold. The only other amino acid occurring more often in thermophiles and hyperthermophiles was tyrosine (Y, [1] Fig. 3.2A). On the other hand, the hydrophobic methionine (M, [1], Fig. 3.2A) was over-represented in both psychrophiles and psychrotolerants. When grouping all amino acids in two classes (hydrophobic/not) using different hydrophobicity scales (Eisenberg and Weiss [120], Kyte-Doolittle [121], and Janin [122]), we could confirm the observation [50]: psychrophiles have less hydrophobic residues than hyperthermophiles (but not less than thermophiles). The differences observed between the opposites (cold/heat, [1],

Fig. 3.2B) were insignificant (Z-score between -0.05 and -0.1- for the psychrophiles vs. 0.04-0.2 for the hyperthermophiles [1]).

Overall, it seemed likely that the differences in disorder content between two very different extreme temperature organisms such as Colwellia (psychrophile) and Pyrococcus (thermophile) are largely attributable to homologous proteins that kept their overall structure with some small alterations to adapt to extreme climates. These alterations may include shorter loops, less surface area and more compact proteins in thermophiles, and exceptionally flexible proteins in psychrophiles.

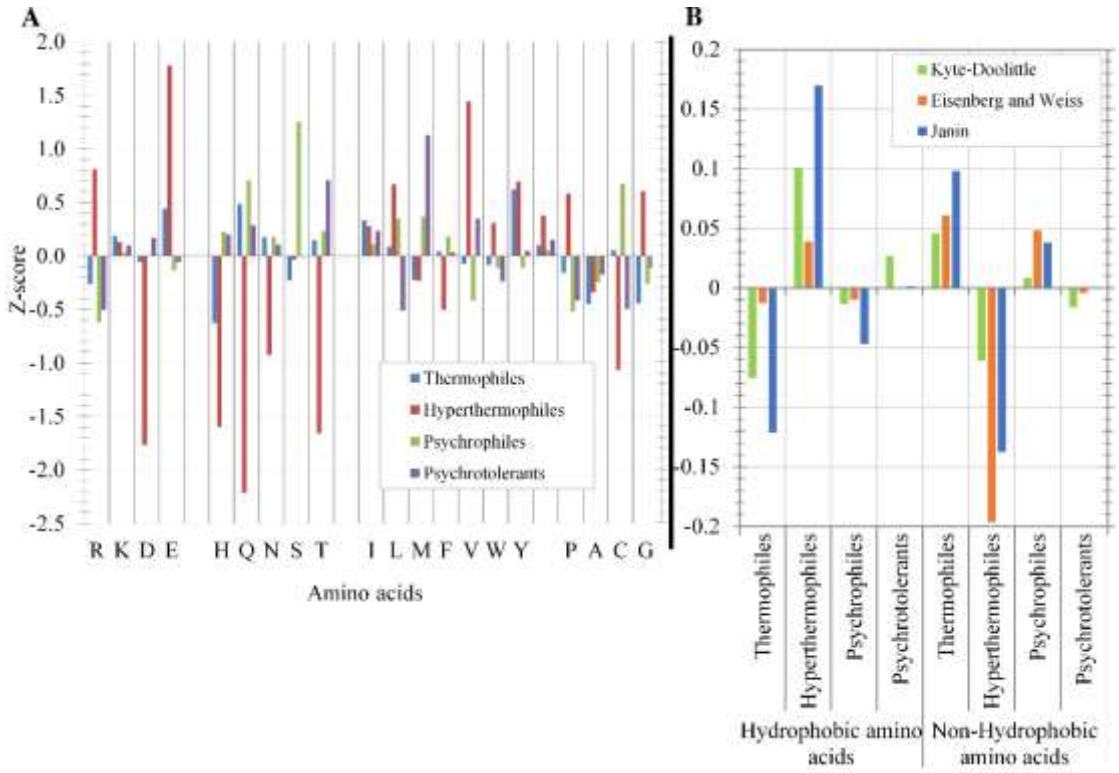

Figure 3.2: Graphical representation of amino acid abundance in different extreme organisms using Z-score [1]**. (A)** The diagram shows the single amino acid abundance in whole proteomes of hyperthermophiles (red), thermophiles (blue), psychrophiles (green) and psychrotolerants (purple) organisms. **(B)** To have a general view of the amino acids content in organisms in terms of hydrophobicity, the amino acids where grouped according to different scales.

3.1.7 Statistical analysis of disorder content distribution: habitat vs. phyla

By applying the Kruskal-Wallis and the paired Wilcoxon-Test on the organisms grouped first into habitat (Table 2.1) and later into phyla (Table 2.3), we found that the groups in habitat presented different distribution of disordered content for MD ($P<0.05$; Fig. 3.3A) and IUPred predictions ($P<0.05$, $P<0.005$; Fig. 3.3B) and for all thresholds (%long30, %long50 [1] and %long80 [1]; Fig. 3.3). Contrarily, the phyla groups in most instances did not differ in any statistically significant way ([1]; Fig. 3.3). Exceptions were found for NORSnet ("loopy" disorder) for all thresholds, but for MD only for the middle long disordered proteins (%long50 [1]) and for IUPred only for the proteins containing long disordered regions (%long80 [1]). Thus, the "loopy" disorder appeared more conserved than other disordered regions [123].

We did not find any convincing explanation for the changes in regions longer than 80 residues, but we observed that other studies support the opposite [32, 124-127]. For the analysis of the completely disordered proteins [1] we found that both, phyla and habitat, influenced the disorder content distribution for the IUPred and NORSnet predictions, but only for disordered regions with lengths of at least 50 consecutives disordered residues (%long50 [1] and %long80 [1]). All these observations were confirmed also using the Z-scores values instead of the raw values [1].

The habitat is a complex reality defined by a variety of factors such as temperature, pH, energy source and metabolism [1]. Therefore, we tried to analyze these factors separately. In doing so we found a significant difference in disorder content between the organisms grouped by temperature (high temperature – low disorder [1]) and by oxygen requirement [1] (an aerobic lifestyle implied higher disorder [128, 129]). For the other factors (metabolism, energy source, cell shape [1]), we did not observe a significant influence on disorder (content of proteins with long disordered regions). Definitely, we could only propose that practically the protein disorder abundance in proteomes was more related to environment than to phylogeny but this might be the inverse for "loopy" disorder.

3.1.8 Null hypothesis rejected: disorder is similar between habitats

From several analyses we already know, protein disorder is much more abundant in eukaryotes than in prokaryotes ([9, 130]). Nevertheless, there are considerable variations between prokaryotes [1] which seem to be more related to habitats than to phyla, *i.e.* proteins from organisms living in similar habitats contains a similar percentage of proteins with long disordered regions, but not proteins with similar phylogeny ([1]; Fig. 3.3). We described some examples for strong correlation between habitat and disorder and also found many examples of organisms for which we observed the correlation contrary to what we expected. Those contradictory observations could be caused by mistakes in the method as the applied prediction methods have not been implemented for the studied type of organisms (organisms

living in extreme habitats). But we don't have any hard evidence to support this observation. For instance, secondary structure prediction methods developed over 20 years ago ([131]) continue to correctly capture the structure for very different proteins from very different environments before they were experimentally observed (disorder just being one case in point [9]). Furthermore, none of the used methods appear to have been developed in any way on data specific to non-extremophiles.

Another point is the different prediction values observed between the three methods used in this study. These differences could be explained easily as they capture different aspects of disorder. Given the heterogeneity of the phenomenon protein disorder, differences between two data sets captured by one method and not by two others may point to the exact reason why that 'outlier' method correctly captures a reality missed by the other two. In my example, we found that the IUPred predictions for radiation resistant organisms/proteins correlate with high disorder which might be more helpful than the MD predictions.

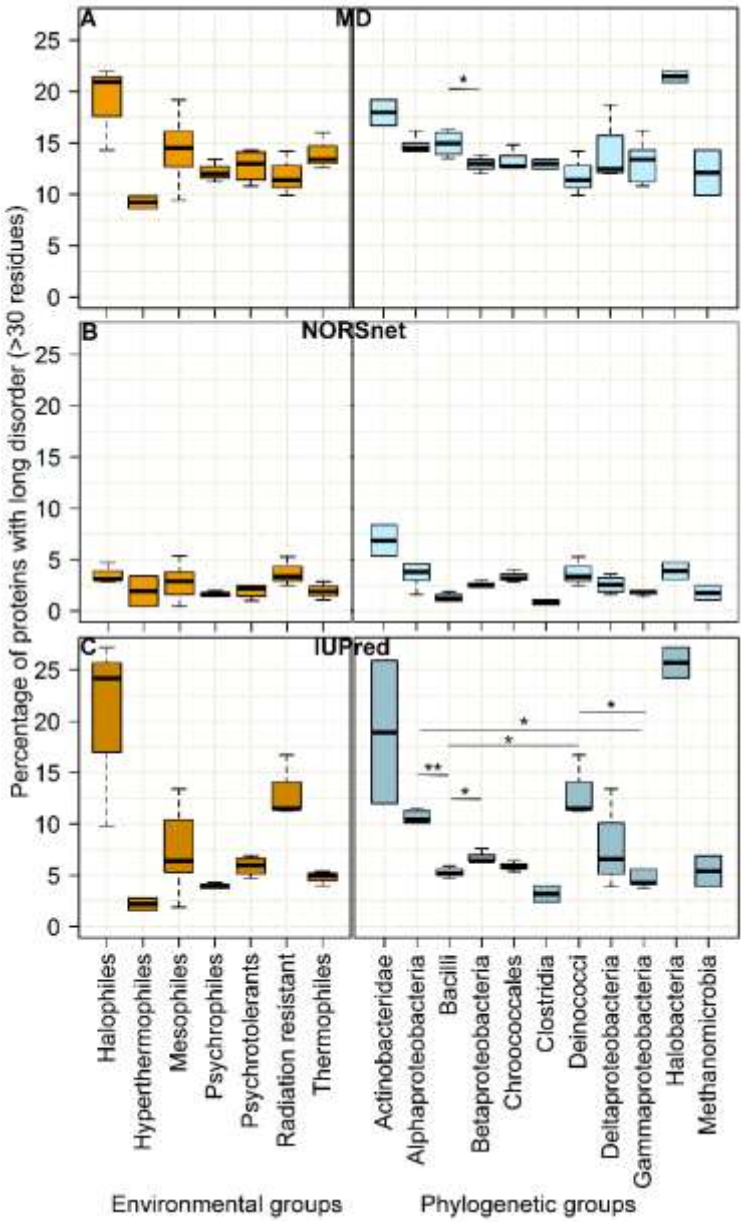

Figure 3.3: Protein disorder content differs for habitat, not for phyla [1]**.** We represent the protein disorder content for the organisms in similar habitats (left panel) and those in the same phyla (right panel). The y-axes give the percentage of proteins with at least one region of ≥30 consecutive residues predicted as disordered by MD (A), NORSnet (B) and IUPred (C). The x-axis on the left side marks the different environmental groups [1]); on the right side, the studied phylogenetic groups [1]). The groups which are significant for a paired Wilcoxon test were marked with * ($P<0.05$) or ** ($P<0.005$).

3.2 Protein disorder content under high temperature pressure

In this part of the thesis we try to find an answer to the specifically duplication of chromosome III and two fragments of chromosomes IV and XII when a culture of Saccharomyces cerevisiae (baker's yeast) survives sudden heat shock. This solution is "temporary" in the sense that is realized "only" for the first 400-1200 generations of the surviving cells; after some 1200 generations more detailed solutions are found.

3.2.1 Duplications reduce the overall amount of protein disorder

Reacting to high temperature, yeast (*S. cerevisiae*) duplicates chromosome III and fragments from chromosomes IV and XII [71] (for short, we use chr. N to point the yeast chromosome N). The 16 yeast chromosomes differ more than 6-fold in their size, but have proteins of similar length [2] (Fig. 2.1). The duplicated chr. III is the third smallest with 183 genes, of which 153 are mapped and 132 constitute "verified ORFs" (smaller are only chr. I and chr. VI with 90 and 125 proteins, respectively [2]). Chr. III, due its small size, was the first complete synthesized functional yeast chromosome [132]. In opposition to protein length, the percentage of proteins with long regions of disorder diverged notably between the 16 yeast chromosomes [2] (Fig. 3.4).

In our analysis, we found that chr. III and chr. X were the chromosomes with the least content of proteins with predicted disorder [2] (Fig. 3.4). It seems, heat response implies a duplication of one of the two chromosomes with the least disorder. Also, for the fragments of chr. IV and chr. XII that were duplicated along with the entire chr. III, presented clearly less disorder than the chromosomes from which they were taken [2] (Fig. 3.4). All of that was consistent with the concept of reducing protein disorder in the response to high temperatures.

Chr. X was the second chromosome with the least content of proteins with predicted disorder, but in contrast to chr. III it was not duplicated. The reason could be in its size, chr. X is more than twice the size of chr. III [2] (Fig. 2.1). The duplication of chr. X implies over twice the cost and that might be too expensive in stress situations such as high temperature. Another explanation could be found through the cellular activities transcribed by chr. X, which may be unimportant for dealing with high temperature. Furthermore, the duplication of chr. X would reduce - due its length (double than chr. III) - the overall protein disorder even more than that of chr. III [2].

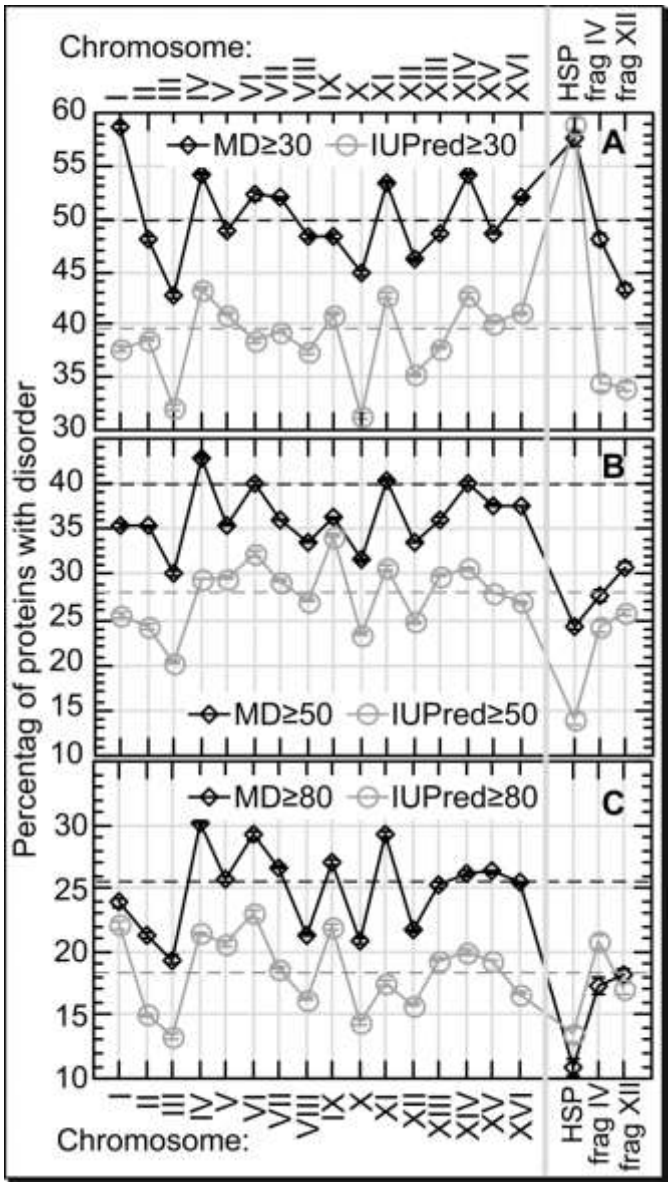

Figure 3.4: Protein disorder differs between yeast chromosomes [2]**.** The composition of proteins with long regions of disorder (y-axes) differed significantly between the 16 chromosomes of ***S. cerevisiae*** (x-axes) and also for the set of HSPs. The three rightmost marks on the x-axes describe: HSPs and the disorder predictions for the HSR-related duplicated fragments on chromosome IV and chromosome XII (frag IV and frag XII). The differences were similar for two different prediction methods (MD in black, IUPred in light gray), and for different thresholds with respect to the minimal length of a disordered region (**A:** ≥30 consecutive residues predicted in disorder, **B**: ≥50, **C**: ≥80). Dashed horizontal lines mark the averages over all chromosomes. Error bars are too small to become visible on the scale chosen. The least disorder content was predicted for chromosome III and chromosome X.

We searched though the genome of yeast for a continuous stretch (within a chromosome) that contained 153 proteins with less disorder than those on chr. III. Our result emphasized the particular character of chr. III [2] (Fig. 3.5): only 3% of all continuous genome fragments with 153 proteins presented less disorder content than chr. III (analogous numbers for chr. X [2]: 5%, 29-protein fragment from chr. IV: 52%, 64-protein fragment from chr. XII: 10%; Fig. 3.5) [2]. These numbers showed that the duplication of chr. III could be THE optimal solution for duplicating 153 proteins with as little disorder as possible in the entire genome of yeast.

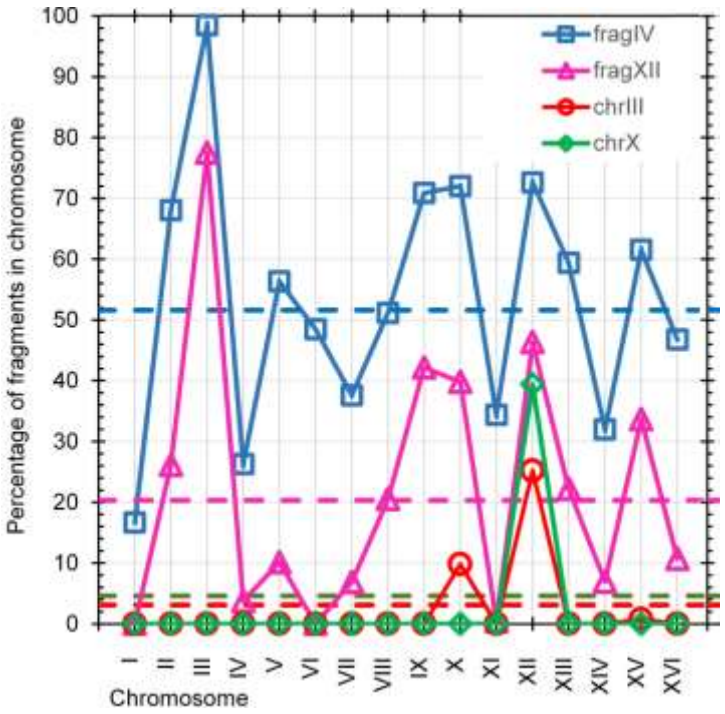

Figure 3.5: Fragments with less disorder than those duplicated during heat shock response (HSR) [2]. We estimated the disorder content in these sets of the duplicated proteins: 55% in proteins of fragment of chromosome IV (blue dashed line), 47% in proteins of fragment of chromosome XII (orange dashed line) and 45% proteins of chromosome III (red dashed line). We also estimated the disorder content for all 347 proteins of chromosome X, which is the non-duplicated chromosome with less disorder content and its proteins are known to not being involved in HSR. The disorder content was 48% (green dashed line) for proteins of chr. X. Then, we screened all chromosomes for stretches encoding 153, 29, 64 and 347 proteins to measure their disorder content and compare it to the disorder content in chromosomes III, X and the fragments of chromosomes IV and XII. For example, chromosome II contains 388 proteins, so we measured the disorder content in 325 stretches encoding 64 proteins (as this is the number of HSR proteins encoded by chromosome IV) to compare their disorder content to that in proteins of chromosome IV. We found that proteins encoded by 68% of regions of chromosome IV have less disorder than 64 proteins of chromosome IV.

3.2.2 Heat-shock proteins

Our model might explain why the duplication of chr. III is better than other duplications. Nevertheless, to answer the question about the selective advantages of duplicating the proteins on chr. III, we assumed that chr. III contained proteins that actively help with the adaptation to heat stress. The instant suspects were heat-shock proteins (HSPs) and proteins known to interact with these (HSP-binders). However, we found the known HSPs and HSP-binders to be distributed over all 16 yeast chromosomes [2] (Fig. 3.6) without any special predilection to chr. III. Furthermore, all regions duplicated in response to heat shock contain only one known HSP (HSP30) and one known HSP-binder (TAH1). This implied that 1.3% of all known HSPs and HSP-binders were duplicated in an event that duplicated 0.5% of all genes [2]. Numerically, this constituted a 2.6 fold over-representation, but statistically, this finding appeared insignificant: less than 1 in 50 of all HSPs and HSP-binders might be convincing if HSP30 and TAH1/HSP90 were THE most important proteins for the given conditions, but they are not [2]. Experimental evidences support these data, introducing an extra copy of HSP30 into wild-type cells did not increment the capacity of the cells to deal with high temperature (Dahan & Pilpel, unpublished).

The known HSPs (Fig. 3.6) marginally altered expression levels in response to heat stress in the course of the fixation of the trisomy but almost all HSPs were notably up-regulated[2] (arrows in Fig. 3.6) when the "refined descendants" replaced the trisomy [71]. This could suggest that the duplicated genes are fundamental for survival under heat shock.

Additionally, HSPs appeared especially plenty of disordered regions of length 30-50 consecutive residues (in particular for IUPred [2]). This observed disorder has already been mentioned in earlier works to be necessary for HSP function [133]. On the other hand, HSPs appear to be limited in longer disorder [2] (>50).

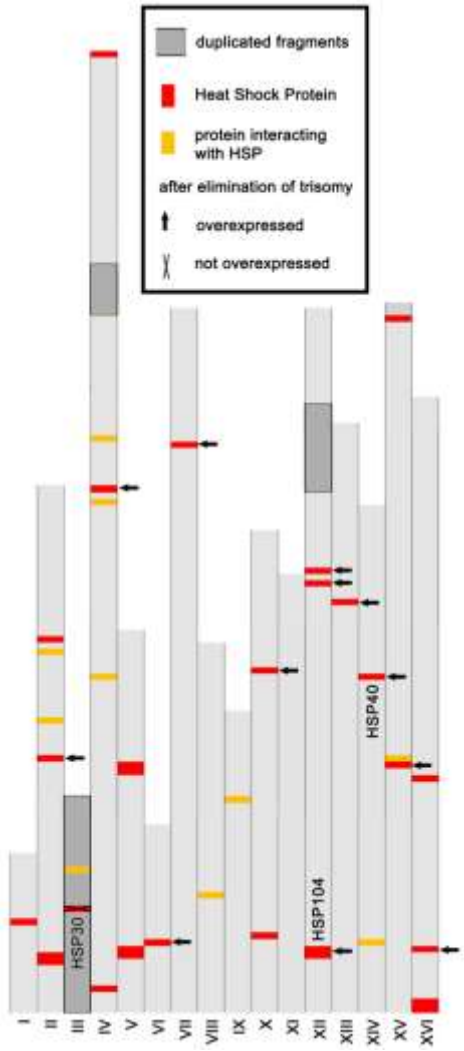

Figure 3.6 Distribution of Heat Shock proteins (HSPs) along chromosomes [2]. The duplicated chr. III and the duplicated fragments of chr. IV (IV) and XII (XII) (dark grey) contain only one known heat shock protein (HSP; on chromosome III, marked in red: HSP30) and one protein known to interact with it (also on chr. III: marked in yellow: TAH1). Overall, known HSPs (marked with crosses and arrows) change their expression levels modestly during the fixation of the trisomy.

3.2.3 GO terms enriched for growth and reproduction in heat stress-duplicate regions

HSPs did not clarify why chr. III was selected to be duplicated in response to heat stress. The only one reason that we found until now for duplication was the protein disorder reduction[2]. But when we went further and tried to find an explanation based on the importance of the proteins on chr. III for growth under high temperature, using a simple scanning of the annotations of, *e.g.* GO [134] annotations, was still not enough: the question was not whether proteins on chr. III had certain functions, but whether these were significantly enough overrepresented to explain why chr. III and not chr. VI or chr. I (the two other small chromosomes) was duplicated in response to high temperatures. To realize such a differential view, we analyzed the GO term enrichment for the duplicated chromosome and chromosomal regions [135].

Growth and reproduction are the most relevant cell activities. The organism still has to grow and proliferate, including extreme conditions. The GO enrichment analysis supported this assumption [2] (Fig. 3.7): the two most prolific GO terms in the heat stress-duplicated regions (entire chr. III + fragment of chr. IV) were those connected to (i) sexual reproduction (Fig. 3.8; "conjugation with cellular fusion", "reproductive cellular process" and "response to pheromone") and to (ii) sugar transport [2] (hexose transport process as well as mannose, fructose and glucose transmembrane transporter activity).

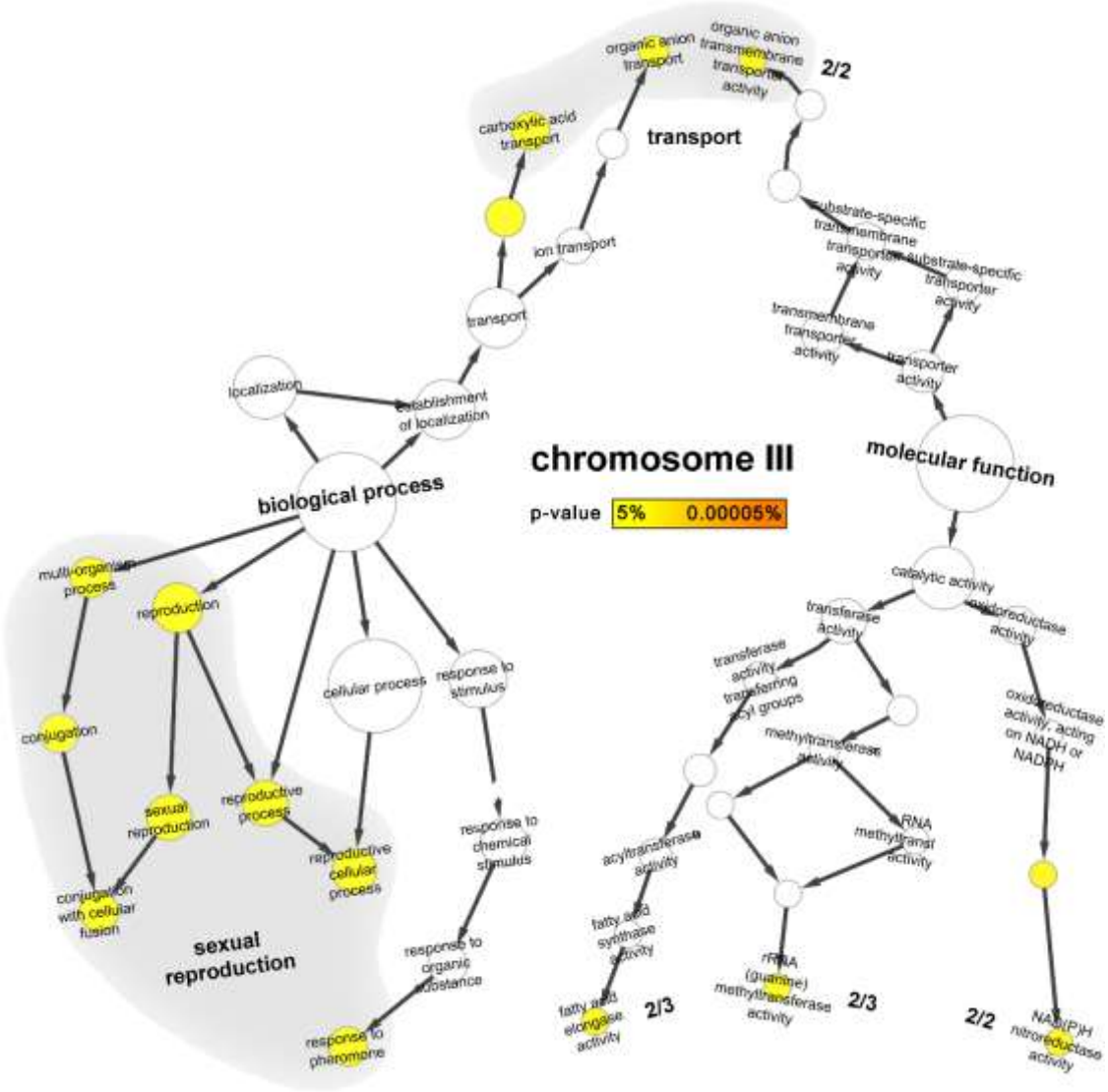

Figure 3.7: GO enrichment of sexual reproduction and nutrient uptake [2]. Depicted are the GO terms for chr. III. The tree gave the complete set of all experimentally annotated GO-terms (Gene Ontology [92]) for any of the proteins on chr. III that describe *biological process* (left branch) and *molecular function* (right branch). The enrichment analysis [135] described how much chr. III GO-terms are enriched with respect to all other GO-terms from yeast: all terms marked by yellow circles were significantly enriched.

The major energy source of yeast is sugar, in particular hexose monosaccharides (C6H12O6; e.g. glucose, fructose, mannose) which are imported into the cell through hexose transporters. These membrane proteins are encoded by *HXT* genes [136, 137] which were almost 5 times over-represented on the duplicated fragment of chr. IV (*HXT3*, *HXT6* and *HXT7*) with respect to random [2]. It is worth pointing out that, several works had encountered duplication of two of these genes (*HXT6* and *HXT7*) in yeast populations evolving under low-nutrient conditions [59, 138].

Sexual reproduction is also fundamental to survival of yeast cultured under heat stress [139, 140]. Seven of the ten processes which were suggested by a standard GO-term enrichment analysis [135] to be significantly overrepresented in the heat stress-duplicated chr. III [2], were involved in reproduction. Three of these processes were related specifically to sexual reproduction [2]; the others pertained to general reproductive processes [2] (Fig. 3.7). In particular, the reproduction-related processes involved cell fusion, pheromone response, nuclear fusion, chromosome disjunction, nuclear segregation after mating, fusion of haploid nuclei during mating, cytokinesis (division of cytoplasm and plasma membrane of a cell and its separation into two daughter cells which is also relevant for asexual mitotic and in the developmental process in which the size of a cell is generated and organized [2]. All The genes involved in these overrepresented functions were also required for the correct localization of other proteins involved in cytokinesis and bud site selection [141-145].

Other important processes and activities overrepresented on chr. III were related to the avoidance of oxidative stress [2] (e.g. carboxylic acid transport, Fig. 3.7); which may be important for the survival since during the vegetative asexual reproduction, cells were exposed to oxidative stress) and NAD(P)H nitro-reductase activity [2] (Fig. 3.7 ; the only nitroreductase related proteins in yeast – HBN1 and FRM2 [146] – were found on chr. III [2]. The proteins involved in these two activities were also implicated in cellular detoxification [147], another task relevant for survival under stress.

Our data defended the point of view that chr. III was essential for sexual reproduction in yeast. However, the laboratory strains of yeast survived the heat stress through asexual reproduction [71], *i.e.* apparently yeast did not need what was so uniquely enriched in the heat stress duplicated chromosome for their survival. The group of proteins recognized to be involved in reproduction on chr. III [2] contains more disordered regions than the average for chr. III [2]. Some of these proteins with long disordered regions might not be functional in heat. As we haven't identified the reasons for the duplication of these duplicated proteins involved in reproduction, we proposed two speculations: First, the genes involved in sexual reproduction might include another cellular activity that was more relevant to the growth conditions applied during the lab evolution experiment [2]. Among other genes, *CDC10* was also required to maintain cell polarity (GO:0030011) and *BUD3* and *BUD5* were involved in axial cellular bud-site selection (GO:0007120). All these activities were also involved in asexual reproduction. Our second suggestion appears to be more labored, more specifically that the disorder-rich proteins associated to sexual reproduction

might have been duplicated accidentally, *i.e.* because they were located on chr. III but not due their significance for the survival in heat [2].

We found other marginally overrepresented processes in the duplicated fragments with some importance for yeast subsistence in heat [2] (fatty acid elongase, , rRNA (guanine) methyltransferase, and the importin-alpha export receptor activities). We investigated these in detail but none of those gave a coherent interpretation.

3.2.4 Protein localization and disorder content

The partial experimental annotation available is one of the most important limitations on functional enrichment studies even for yeast, an organism profoundly studied and analyzed in the laboratory. May be all our above explained speculations omit the real motivation; it is possible that the function of the really important proteins are not yet identified. Therefore, we extended our analysis with the prediction of sub-cellular localization of all yeast proteins. Although, the experimental localization annotations for yeast are still partial and covering only about 70% of all proteins [148], there are very reliable prediction methods, such as LocTree3 [148] which can make decisive differences when comparing 'complete' data sets [149]. In our study, we found nuclear proteins to be undoubtedly scarce on chr. III [2](-4.6 percentage points with respect to the entire proteome, Fig. 3.8A). In the other hand, secreted (extra-cellular) or annotated proteins as endoplasmic reticulum (ER) or membrane proteins (each 3.2 percentage points higher than in the full yeast proteome) were found abundantly in chr. III [2]. We also observed significantly more disorder in nuclear proteins [2] (nuclear 77% *vs.* <40% for non-nuclear, Fig. 3.8) which may explain the nuclear proteins shortage on chr. III.

Although these data is clear, the interpretation unfortunately is not easy. The abundance of secreted proteins on chr. III [2] (about 3.2 percentage points more on chr. III than in entire yeast, Fig. 3.8A) may suggest that more proteins are secreted into the 'hot' environment to deal with the heat shock. Based on the correlation between habitat and disorder [150], we suppose that proteins are more probable to withstand high temperatures but with less disorder. Unluckily, we did not get any convincing or sufficiently clear answers by applying a GO enrichment study to the secreted proteins ([2]; Fig: 3.8A).

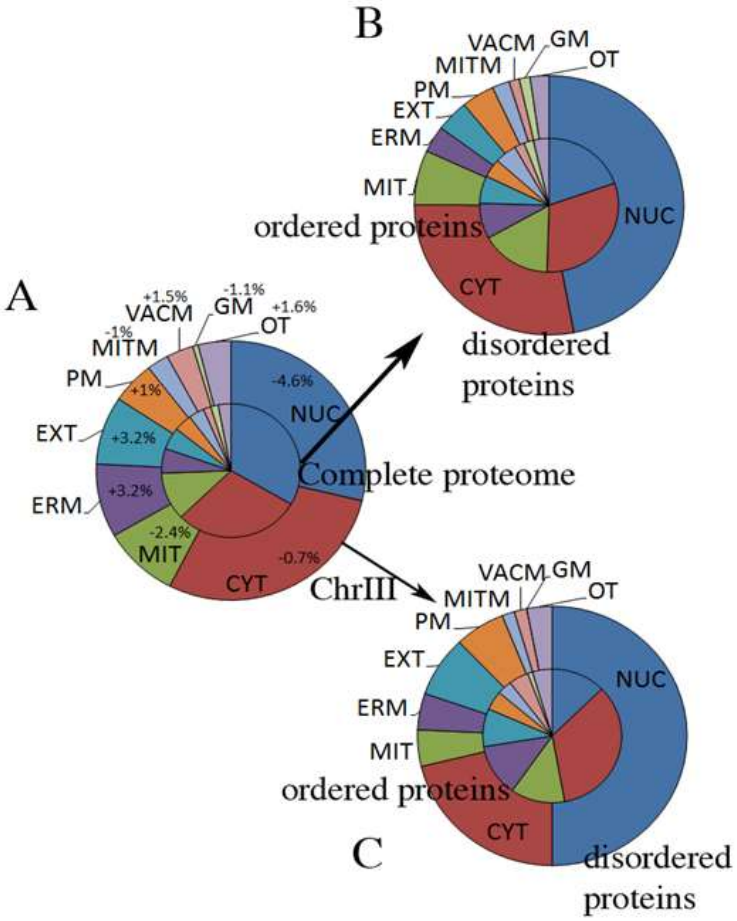

Figure 3.8: Distribution comparison of chr. III proteins and the complete yeast proteome across localization classes [2]. **(A)** Comparison of the distribution across the compartments for all proteins in yeast (inner pie, 5667 proteins) and for proteins of chromosome III (outer ring, 153 proteins). For each compartment we provide the difference in two distributions. **(B)** Comparison of the distribution of predicted (method %long30 and MD) ordered (inner pie) and disordered proteins (outer ring) across different localization compartments in the entire proteome of yeast and **(C)** in the proteins encoded by chromosome III. Abbreviations: NUC, nucleus; CYT, cytoplasm; MIT, mitochondria; ERM, Endoplasmic Reticulum membrane; EXT, extra-cellular; PM, plasma membrane; MITM: mitochondria membrane; VACM, vacuole membrane; GM, Golgi apparatus membrane; OT, other (including Golgi apparatus, Endoplasmic Reticulum, vacuole, peroxisome, peroxisome membrane and nucleus membrane.

3.2.5 Protein-Protein Interactions (PPI)

The majority of proteins does not work alone and are interacting with other proteins for a proper function. To fully understand their function and primary importance in the wide range of protein interactions pathways and complete the general study about structure, functionality and localization of the yeast proteins, we should considerer

them in the context of their interacting partners and not as isolated objects in the cells. Therefore, we compared the network of experimentally characterized protein-protein interactions (PPIs) between the entire yeast and those fragments that are duplicated in heat evolving populations.

The degree (number of interactions per protein) was significantly lower for the duplicated chr. III [2] (average=16±2, Fig. 3.9A). Similarly, trend was found for the betweenness (number of times that a protein acts as a bridge along the shortest path between two other proteins: average=1800±300 Fig. 3.9B). Furthermore, chr. III is one of the chromosomes with the largest mean value for the average degree for their neighbors (average=380±40; Fig. 3.9C). These network analyses may suggest that chr. III might also be a good choice for a first line of defense against high temperature because the proteins encoded on this chromosome play less essential roles for the overall PPI network. Neither the chr. III proteins themselves nor their PPI neighbors tend to be hubs.

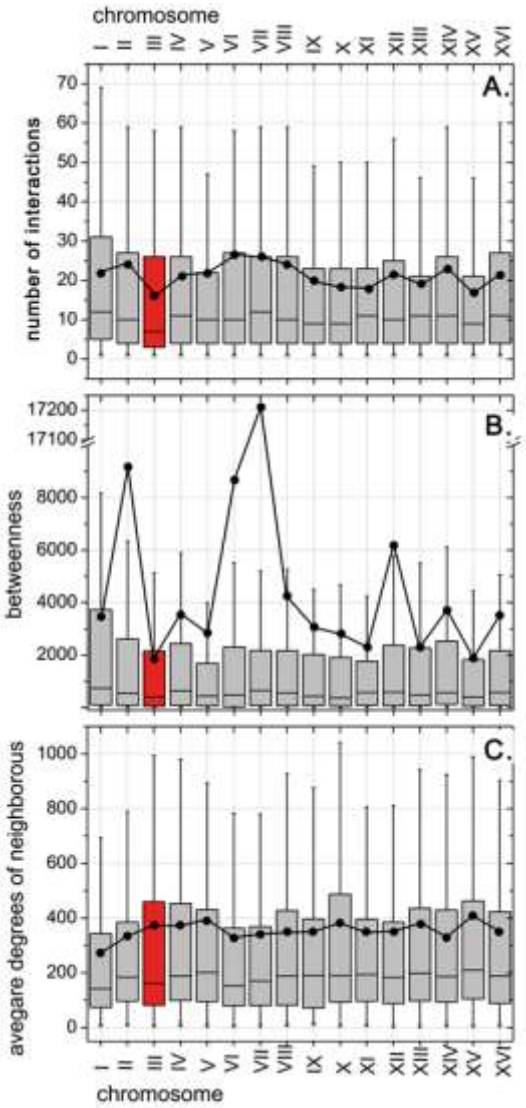

Figure 3.9: Protein-protein interaction (PPI) network differs between yeast chromosomes. Plotted for each yeast chromosome is: (A) Degree, The number of PPIs per protein. The chr. III is highlighted in red; the median is the black dot in the box. (**B**) Betweenness, number of times that a node acts as a bridge along the shortest path between two other nodes and (**C**) Average neighbor degree

4 Conclusion

The purpose of this thesis was the study of intrinsically disorder proteins in terms of extreme environments. For that we analyzed first the influence of the habitat on disorder content in extremophiles, organisms living in environments with extreme conditions. Extremophiles thrive in environments with extreme conditions such as high salt, exceptionally low or high temperatures and high radiation. We compared organisms through a quite simple criterion, namely the percentage of proteins for which at least one long region of disorder was predicted by two complementary approaches. We analyzed protein disorder for several prokaryotic extremophiles and their closest phylogenetic relatives. We found protein disorder to be more reflective of habitat than of the evolutionary relation. This suggested that disordered regions might crucially help in adapting to challenging environments.

For example, halophiles presented significantly more protein disorder than their mesophilic relatives suggesting that protein disorder may compensate for the osmotic stress in extremely salty environments. Furthermore, we showed that the differences in the disorder abundance among organisms from different habitats were independent of their corresponding taxonomic branch by using a large set of proteomes representing different branches of the tree of life. For instance, both halophilic bacterial and halophilic archaeal proteomes were more disordered than their taxonomic neighbors. Moreover, the hyperthermophile organisms appeared to have fewer disordered proteins than their mesophilic taxonomic relatives to compensate for the impact on their 3D structures.

Finally, we investigated how disordered regions might contribute to environmental adaptation. Comparing the homologues between two extremophiles from cold and heat, we established that more often than expected by chance, disordered regions were found in the cold than in the heat. Largely, it appeared that the level of disorder was rather affected by many small changes than by few big ones. Clearly, we once again established protein disorder as an important building block to bring about evolutionary changes such as the adaptation to different habitats.

In the second part of our study we refined our target and centered the study on only one organism (*Saccharomyces cerevisiae*) and one specific sudden environmental change (heat shock). Experimentally it has already been proved that organisms can duplicate their whole genome or particular chromosomes (aneuploidy) in response to sudden dramatic changes in the environment. As such broad changes are costly, using aneuploidy gives way to more specialized, focused solutions that require many generations to evolve.

The entire chromosome III and two fragments from chromosomes IV and XII in a culture of budding yeast were duplicated as a “transient evolutionary solution” in response to high temperature - a “transition” that fostered the survival of between 400 and 2,000 generations. Here, we reported that while the proteins on all 16 main

chromosomes from yeast have similar length, they differ substantially in terms of the fraction of long regions predicted to contain protein disorder (≥30-80 consecutive residues predicted as disordered by IUPred and MD). We found the regions duplicated under heat stress depleted of predicted disorder. In fact, chromosome III was one of the two chromosomes with the least disorder. The other (chromosome X) is twice as large, i.e. would cost significantly more to duplicate. Decreasing the overall content in protein disorder is likely an important strategy to protect against heat stress. A detailed analysis of the experimentally characterized protein-protein interaction (PPI) network in yeast revealed the duplicated proteins to be connected less than average which supported why the duplicated regions might not cause damage and therefore could be used as a fast and reliable solution to lead with hot stress.

When studying the advantages of duplicating exactly these regions we found no sustained evidence for an over-representation of heat-shock proteins (HSPs) in the duplication. Instead, a Gene Ontology (GO) enrichment analysis suggested that the duplicated regions were enriched in processes related to reproduction and to the import of nutrients. The set of GO enriched proteins appeared so important that they were duplicated although high in disorder. This might point to where the explanation for the duplication might be found. Overall, our data suggested a very simple hypothesis: identify the region with lowest protein disorder that is large enough, not too large and duplicate it along with possibly other fragments that are also depleted of disorder in order to cope with heat stress.

Bibliography

1. Vicedo E, Schlessinger A, Rost B. Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes. PLoS One. 2015;10(8):e0133990. doi: 10.1371/journal.pone.0133990. PubMed PMID: 26252577; PubMed Central PMCID: PMC4529154.
2. Esmeralda Vicedo ZG, Yu-An Dong, Tatyana Goldberg, Burkhard Rost. Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock F1000Research. 2015;4(1222). doi: 10.12688/f1000research.7178.1.
3. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder--a breakthrough invention of evolution? Curr Opin Struct Biol. 2011;21(3):412-8. Epub 2011/04/26. doi: S0959-440X(11)00066-2 [pii]

10.1016/j.sbi.2011.03.014. PubMed PMID: 21514145.
4. Anfinsen CB, Scheraga HA. Experimental and theoretical aspects of protein folding. Adv Prot Chem. 1975;29:205-300.
5. Karplus M, Weaver DL. Protein folding dynamics. Nature. 1976;260:404-6.
6. Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976;261:552-8.
7. Levitt M, Warshel A. Computer simulation of protein folding. Nature. 1975;253:694-8.
8. Wright PE, Dyson HJ. Linking folding and binding. Curr Opin Struct Biol. 2009;19(1):31-8. Epub 2009/01/23. doi: S0959-440X(08)00179-6 [pii]

10.1016/j.sbi.2008.12.003. PubMed PMID: 19157855; PubMed Central PMCID: PMC2675572.
9. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol. 2002;322(1):53-64. Epub 2002/09/07. doi: S0022283602007362 [pii]. PubMed PMID: 12215414.
10. Devos D, Dokudovskaya S, Williams R, Alber F, Eswar N, Chait BT, et al. Simple fold composition and modular architecture of the nuclear pore complex. Proc Natl Acad Sci U S A. 2006;103(7):2172-7. Epub 2006/02/08. doi: 0506345103 [pii]

10.1073/pnas.0506345103. PubMed PMID: 16461911; PubMed Central PMCID: PMC1413685.
11. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005;272(20):5129-48. Epub 2005/10/13. doi: EJB4948 [pii]

10.1111/j.1742-4658.2005.04948.x. PubMed PMID: 16218947.
12. Kosol S, Contreras-Martos S, Cedeno C, Tompa P. Structural characterization of intrinsically disordered proteins by NMR spectroscopy. Molecules. 2013;18(9):10802-28. doi: 10.3390/molecules180910802. PubMed PMID: 24008243.
13. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J. 2007;92(5):1439-56. PubMed PMID: 17158572.

14. Dunker AK, Gough J. Sequences and topology: intrinsic disorder in the evolving universe of protein structure. Curr Opin Struct Biol. 2011;21(3):379-81. Epub 2011/05/03. doi: 10.1016/j.sbi.2011.04.002. PubMed PMID: 21530236.
15. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008;18(6):756-64. Epub 2008/10/28. doi: S0959-440X(08)00151-6 [pii]

10.1016/j.sbi.2008.10.002. PubMed PMID: 18952168.
16. Greenleaf WJ, Woodside MT, Block SM. High-resolution, single-molecule measurements of biomolecular motion. Annual review of biophysics and biomolecular structure. 2007;36:171-90. doi: 10.1146/annurev.biophys.36.101106.101451. PubMed PMID: 17328679; PubMed Central PMCID: PMC1945240.
17. Roy R, Hohng S, Ha T. A practical guide to single-molecule FRET. Nature methods. 2008;5(6):507-16. doi: 10.1038/nmeth.1208. PubMed PMID: 18511918; PubMed Central PMCID: PMC3769523.
18. Schuler B, Eaton WA. Protein folding studied by single-molecule FRET. Curr Opin Struct Biol. 2008;18(1):16-26. doi: 10.1016/j.sbi.2007.12.003. PubMed PMID: 18221865; PubMed Central PMCID: PMC2323684.
19. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. PLoS Comput Biol. 2007;3(7):e140. Epub 2007/07/31. doi: 06-PLCB-RA-0416 [pii]

10.1371/journal.pcbi.0030140. PubMed PMID: 17658943; PubMed Central PMCID: PMC1924875.
20. Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics. 2007;23(18):2376-84. Epub 2007/08/22. doi: btm349 [pii]

10.1093/bioinformatics/btm349. PubMed PMID: 17709338.
21. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. PLoS One. 2009;4(2):e4433. Epub 2009/02/12. doi: 10.1371/journal.pone.0004433. PubMed PMID: 19209228; PubMed Central PMCID: PMC2635965.
22. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005;21(16):3433-4. Epub 2005/06/16. doi: bti541 [pii]

10.1093/bioinformatics/bti541. PubMed PMID: 15955779.
23. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005;347(4):827-39. Epub 2005/03/17. doi: S0022-2836(05)00129-4 [pii]

10.1016/j.jmb.2005.01.071. PubMed PMID: 15769473.
24. Burra PV, Kalmar L, Tompa P. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PLoS One. 2010;5(8):e12069. doi: 10.1371/journal.pone.0012069. PubMed PMID: 20711457; PubMed Central PMCID: PMC2920320.

25. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SC. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics. 2014. doi: 10.1093/bioinformatics/btu625. PubMed PMID: 25246432.
26. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71. Epub 2001/11/09. PubMed PMID: 11700597.
27. Esnouf RM, Hamer R, Sussman JL, Silman I, Trudgian D, Yang ZR, et al. Honing the in silico toolkit for detecting protein disorder. Acta Crystallogr D Biol Crystallogr. 2006;62(Pt 10):1260-6. Epub 2006/09/27. doi: S0907444906033580 [pii]

10.1107/S0907444906033580. PubMed PMID: 17001103.
28. Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, et al. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 12(2):R14. Epub 2011/02/18. doi: gb-2011-12-2-r14 [pii]

10.1186/gb-2011-12-2-r14. PubMed PMID: 21324131; PubMed Central PMCID: PMC3188796.
29. Mohan A, Sullivan WJ, Jr., Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol Biosyst. 2008;4(4):328-40. Epub 2008/03/21. doi: 10.1039/b719168e. PubMed PMID: 18354786.
30. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. Biochem Cell Biol. 88(2):167-74. Epub 2010/05/11. doi: o09-163 [pii]

10.1139/o09-163. PubMed PMID: 20453919.
31. Montanari F, Shields DC, Khaldi N. Differences in the number of intrinsically disordered regions between yeast duplicated proteins, and their relationship with functional divergence. PLoS One. 2011;6(9):e24989. doi: 10.1371/journal.pone.0024989. PubMed PMID: 21949823; PubMed Central PMCID: PMC3174238.
32. van der Lee R, Lang B, Kruse K, Gsponer J, Sanchez de Groot N, Huynen MA, et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell reports. 2014;8(6):1832-44. doi: 10.1016/j.celrep.2014.07.055. PubMed PMID: 25220455.
33. Tompa P, Prilusky J, Silman I, Sussman JL. Structural disorder serves as a weak signal for intracellular protein degradation. Proteins. 2008;71(2):903-9. doi: 10.1002/prot.21773. PubMed PMID: 18004785.
34. Gsponer J, Futschik ME, Teichmann SA, Babu MM. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science. 2008;322(5906):1365-8. doi: 10.1126/science.1163581. PubMed PMID: 19039133; PubMed Central PMCID: PMC2803065.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637. doi: 10.1002/bip.360221211. PubMed PMID: 6667333.
36. Zheng WM. Clustering of amino acids for protein secondary structure prediction. J Bioinform Comput Biol. 2004;2(2):333-42. PubMed PMID: 15297985.
37. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002;420(6912):218-23. Epub 2002/11/15. doi: 10.1038/nature01256

nature01256 [pii]. PubMed PMID: 12432406.
38. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9(1):56-68. Epub 1991/01/01. doi: 10.1002/prot.340090107. PubMed PMID: 2017436.
39. Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci. 2001;10(10):1970-9. Epub 2001/09/22. doi: 10.1110/ps.10101. PubMed PMID: 11567088; PubMed Central PMCID: PMC2374214.
40. Aravind L, Koonin EV. Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system. Genome Res. 2000;10(8):1172-84. Epub 2000/08/25. PubMed PMID: 10958635; PubMed Central PMCID: PMC310937.
41. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helix prediction at 95% accuracy. Protein Science. 1995;4:521-33.
42. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proceedings of the National Academy of Sciences. 1997;94(22):11911-6.
43. Rost B. Did evolution leap to create the protein universe? Curr Opin Struct Biol. 2002;12(3):409-16. Epub 2002/07/20. doi: S0959440X02003378 [pii]. PubMed PMID: 12127462.
44. Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, et al. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. PLoS Biol. 2004;2(12):e380. Epub 2004/11/04. doi: 10.1371/journal.pbio.0020380. PubMed PMID: 15523559; PubMed Central PMCID: PMC524472.
45. Petsko GA. Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat". Methods Enzymol. 2001;334:469-78. Epub 2001/06/12. doi: S0076-6879(01)34486-5 [pii]. PubMed PMID: 11398484.
46. Robinson-Rechavi M, Alibes A, Godzik A. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima. J Mol Biol. 2006;356(2):547-57. Epub 2005/12/27. doi: S0022-2836(05)01479-8 [pii]

10.1016/j.jmb.2005.11.065. PubMed PMID: 16375925.
47. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000;13(3):179-91. Epub 2000/04/25. PubMed PMID: 10775659.
48. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics. 2000;1(1):76-88. Epub 2002/01/17. doi: 10.1007/s101420000003. PubMed PMID: 11793224.
49. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. Proteins. 2004;54(1):20-40. Epub 2004/01/06. doi: 10.1002/prot.10559. PubMed PMID: 14705021.
50. D'Amico S, Collins T, Marx JC, Feller G, Gerday C. Psychrophilic microorganisms: challenges for life. EMBO Rep. 2006;7(4):385-9. Epub 2006/04/06. doi: 7400662 [pii]

10.1038/sj.embor.7400662. PubMed PMID: 16585939; PubMed Central PMCID: PMC1456908.
51. Paul S, Bag SK, Das S, Harvill ET, Dutta C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 2008;9(4):R70. Epub 2008/04/10. doi: gb-2008-9-4-r70 [pii]

10.1186/gb-2008-9-4-r70. PubMed PMID: 18397532; PubMed Central PMCID: PMC2643941.
52. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274(5287):546, 63-7. Epub 1996/10/25. PubMed PMID: 8849441.
53. Alberghina L, Mavelli G, Drovandi G, Palumbo P, Pessina S, Tripodi F, et al. Cell growth and cell cycle in Saccharomyces cerevisiae: basic regulatory design and protein-protein interaction network. Biotechnol Adv. 2011;30(1):52-72. Epub 2011/08/09. doi: S0734-9750(11)00120-0 [pii]

10.1016/j.biotechadv.2011.07.010. PubMed PMID: 21821114.
54. Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet. 2002;41(1):1-10. Epub 2002/06/20. doi: 10.1007/s00294-002-0285-z. PubMed PMID: 12073094.
55. Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. Nat Biotechnol. 2004;22(10):1249-52. Epub 2004/10/08. doi: nbt1020 [pii]

10.1038/nbt1020. PubMed PMID: 15470464.
56. Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, Dunham MJ, et al. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science. 2007;317(5840):916-24. Epub 2007/08/19. doi: 317/5840/916 [pii]

10.1126/science.1142210. PubMed PMID: 17702937.
57. Pavelka N, Rancati G, Zhu J, Bradford WD, Saraf A, Florens L, et al. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature. 2010;468(7321):321-5. Epub 2010/10/22. doi: nature09529 [pii]

10.1038/nature09529. PubMed PMID: 20962780; PubMed Central PMCID: PMC2978756.
58. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. Nat Rev Genet. 2012;13(3):189-203. Epub 2012/01/25. doi: nrg3123 [pii]

10.1038/nrg3123. PubMed PMID: 22269907.
59. Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, Ward A, et al. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. PLoS Genet. 2008;4(12):e1000303. Epub 2008/12/17. doi: 10.1371/journal.pgen.1000303. PubMed PMID: 19079573; PubMed Central PMCID: PMC2586090.
60. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, et al. Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet. 2000;25(3):333-7. Epub 2000/07/11. doi: 10.1038/77116. PubMed PMID: 10888885.
61. Rancati G, Pavelka N, Fleharty B, Noll A, Trimble R, Walton K, et al. Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. Cell. 2008;135(5):879-93. Epub 2008/12/02. doi: S0092-8674(08)01196-3 [pii]

10.1016/j.cell.2008.09.039. PubMed PMID: 19041751; PubMed Central PMCID: PMC2776776.
62. Selmecki A, Gerami-Nejad M, Paulson C, Forche A, Berman J. An isochromosome confers drug resistance in vivo by amplification of two genes, ERG11 and TAC1. Mol Microbiol. 2008;68(3):624-41. Epub 2008/03/28. doi: MMI6176 [pii]

10.1111/j.1365-2958.2008.06176.x. PubMed PMID: 18363649.
63. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, et al. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2002;99(25):16144-9. Epub 2002/11/26. doi: 10.1073/pnas.242624799

242624799 [pii]. PubMed PMID: 12446845; PubMed Central PMCID: PMC138579.
64. Polakova S, Blume C, Zarate JA, Mentel M, Jorck-Ramberg D, Stenderup J, et al. Formation of new chromosomes as a virulence mechanism in yeast Candida glabrata. Proc Natl Acad Sci U S A. 2009;106(8):2688-93. Epub 2009/02/11. doi: 0809793106 [pii]

10.1073/pnas.0809793106. PubMed PMID: 19204294; PubMed Central PMCID: PMC2637908.
65. Fawcett JA, Maere S, Van de Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A. 2009;106(14):5737-42. Epub 2009/03/28. doi: 0900906106 [pii]

10.1073/pnas.0900906106. PubMed PMID: 19325131; PubMed Central PMCID: PMC2667025.
66. Torres EM, Williams BR, Amon A. Aneuploidy: cells losing their balance. Genetics. 2008;179(2):737-46. Epub 2008/06/19. doi: 179/2/737 [pii]

10.1534/genetics.108.090878. PubMed PMID: 18558649; PubMed Central PMCID: PMC2429870.
67. Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, Humpton TJ, et al. Aneuploidy drives genomic instability in yeast. Science. 2011;333(6045):1026-30. Epub 2011/08/20. doi: 333/6045/1026 [pii]

10.1126/science.1206412. PubMed PMID: 21852501; PubMed Central PMCID: PMC3278960.
68. Torres EM, Dephoure N, Panneerselvam A, Tucker CM, Whittaker CA, Gygi SP, et al. Identification of aneuploidy-tolerating mutations. Cell. 2010;143(1):71-83. Epub 2010/09/21. doi: S0092-8674(10)01007-X [pii]

10.1016/j.cell.2010.08.038. PubMed PMID: 20850176; PubMed Central PMCID: PMC2993244.
69. Sheltzer JM, Amon A. The aneuploidy paradox: costs and benefits of an incorrect karyotype. Trends Genet. 2011;27(11):446-53. Epub 2011/08/30. doi: S0168-9525(11)00118-1 [pii]

10.1016/j.tig.2011.07.003. PubMed PMID: 21872963; PubMed Central PMCID: PMC3197822.
70. Pavelka N, Rancati G, Li R. Dr Jekyll and Mr Hyde: role of aneuploidy in cellular adaptation and cancer. Curr Opin Cell Biol. 2010;22(6):809-15. Epub 2010/07/27. doi: S0955-0674(10)00099-2 [pii]

10.1016/j.ceb.2010.06.003. PubMed PMID: 20655187; PubMed Central PMCID: PMC2974767.

71. Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, et al. Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A. 2012;109(51):21010-5. Epub 2012/12/01. doi: 1211150109 [pii]

10.1073/pnas.1211150109. PubMed PMID: 23197825; PubMed Central PMCID: PMC3529009.
72. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40(Database issue):D571-9. Epub 2011/12/03. doi: gkr1100 [pii]

10.1093/nar/gkr1100. PubMed PMID: 22135293; PubMed Central PMCID: PMC3245063.
73. Consortium TU. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2011;40(Database issue):D71-5. Epub 2011/11/22. doi: gkr981 [pii]

10.1093/nar/gkr981. PubMed PMID: 22102590; PubMed Central PMCID: PMC3245120.
74. Mancinelli LJRRL. Life in extreme environments. Nature 2001;409:

1092-101. doi: 10.1038/35059215.
75. Morita RY. Biological limits of temperature and pressure. Orig Life. 1980;10(3):215-22. Epub 1980/09/01. PubMed PMID: 7413183.
76. Morita RY. Psychrophilic bacteria. Bacteriol Rev. 1975;39(2):144-67. Epub 1975/06/01. PubMed PMID: 1095004; PubMed Central PMCID: PMC413900.
77. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2011;40(Database issue):D700-5. Epub 2011/11/24. doi: gkr1029 [pii]

10.1093/nar/gkr1029. PubMed PMID: 22110037; PubMed Central PMCID: PMC3245034.
78. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Research. 1997;25:3389-402.
79. Abraham WR, Strompl C, Meyer H, Lindholst S, Moore ER, Christ R, et al. Phylogeny and polyphasic taxonomy of Caulobacter species. Proposal of Maricaulis gen. nov. with Maricaulis maris (Poindexter) comb. nov. as the type species, and emended description of the genera Brevundimonas and Caulobacter. Int J Syst Bacteriol. 1999;49 Pt 3:1053-73. Epub 1999/07/30. PubMed PMID: 10425763.
80. Mika S, Rost B. UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res. 2003;31(13):3789-91. Epub 2003/06/26. PubMed PMID: 12824419; PubMed Central PMCID: PMC169026.
81. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol. 2000;18(6):609-13. Epub 2000/06/03. doi: 10.1038/76443. PubMed PMID: 10835597.
82. Natale DA, Galperin MY, Tatusov RL, Koonin EV. Using the COG database to improve gene recognition in complete genomes. Genetica. 2000;108(1):9-17. Epub 2001/01/06. PubMed PMID: 11145426.

83. Chan Y, Walmsley RP. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. Physical therapy. 1997;77(12):1755-62. PubMed PMID: 9413454.
84. Fan C, Zhang D. A note on power and sample size calculations for the Kruskal-Wallis test for ordered categorical data. Journal of biopharmaceutical statistics. 2012;22(6):1162-73. doi: 10.1080/10543406.2011.578313. PubMed PMID: 23075015.
85. Fitzgerald S, Dimitrov D, Rumrill P. The basics of nonparametric statistics. Work. 2001;16(3):287-92. PubMed PMID: 12441458.
86. Shott S. Nonparametric statistics. Journal of the American Veterinary Medical Association. 1991;198(7):1126-8. PubMed PMID: 2045326.
87. Wolfe MHaDA. Nonparametric Statistical Methods. New York: John Wiley & Sons. 1973:27–33.
88. Wu P, Han Y, Chen T, Tu XM. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. Statistics in medicine. 2014;33(8):1261-71. doi: 10.1002/sim.6026. PubMed PMID: 24132928.
89. Fox DJaW, S. An R and S Plus Companion to Applied Regression. Second Edition S, editor2011.
90. Fox J. Applied Regression Analysis and Generalized Linear Models. Sage SE, editor2008.
91. Team RC. R: A Language and Environment for Statistical Computing. In: Computing RFfS, editor. Vienna, Austria2013.
92. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. Epub 2000/05/10. doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMC3037419.
93. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448-9. Epub 2005/06/24. doi: bti551 [pii]

10.1093/bioinformatics/bti551. PubMed PMID: 15972284.
94. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. Epub 2003/11/05. doi: 10.1101/gr.1239303

13/11/2498 [pii]. PubMed PMID: 14597658; PubMed Central PMCID: PMC403769.
95. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. Stat Methods Med Res. 2008;17(4):347-88. Epub 2007/08/19. doi: 0962280206079046 [pii]

10.1177/0962280206079046. PubMed PMID: 17698936.
96. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist 2001; 29(4):1165-88.
97. Goo YA, Yi EC, Baliga NS, Tao WA, Pan M, Aebersold R, et al. Proteomic analysis of an extreme halophilic archaeon, Halobacterium sp. NRC-1. Mol Cell Proteomics. 2003;2(8):506-24. Epub 2003/07/23. doi: 10.1074/mcp.M300044-MCP200

M300044-MCP200 [pii]. PubMed PMID: 12872007.

98. Oren A, Ginzburg M, Ginzburg BZ, Hochstein LI, Volcani BE. Haloarcula marismortui (Volcani) sp. nov., nom. rev., an extremely halophilic bacterium from the Dead Sea. Int J Syst Bacteriol. 1990;40(2):209-10. Epub 1990/04/01. PubMed PMID: 11536469.
99. Xia Q, Hendrickson EL, Zhang Y, Wang T, Taub F, Moore BC, et al. Quantitative proteomics of the archaeon Methanococcus maripaludis validated by microarray analysis and real time PCR. Mol Cell Proteomics. 2006;5(5):868-81. Epub 2006/02/21. doi: M500369-MCP200 [pii]

10.1074/mcp.M500369-MCP200. PubMed PMID: 16489187; PubMed Central PMCID: PMC2655211.
100. Ohta Y, Hatada Y, Nogi Y, Li Z, Ito S, Horikoshi K. Cloning, expression, and characterization of a glycoside hydrolase family 86 beta-agarase from a deep-sea Microbulbifer-like isolate. Appl Microbiol Biotechnol. 2004;66(3):266-75. Epub 2004/10/19. doi: 10.1007/s00253-004-1757-5. PubMed PMID: 15490156.
101. Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. 2013;22(6):693-724. doi: 10.1002/pro.2261. PubMed PMID: 23553817; PubMed Central PMCID: PMC3690711.
102. Gunbin KV, Afonnikov DA, Kolchanov NA. Molecular evolution of the hyperthermophilic archaea of the Pyrococcus genus: analysis of adaptation to different environmental conditions. BMC Genomics. 2009;10:639. Epub 2010/01/01. doi: 1471-2164-10-639 [pii]

10.1186/1471-2164-10-639. PubMed PMID: 20042074; PubMed Central PMCID: PMC2816203.
103. Sako Y, Nomura N, Uchida A, Ishida Y, Morii H, Koga Y, et al. Aeropyrum pernix gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100 degrees C. Int J Syst Bacteriol. 1996;46(4):1070-7. Epub 1996/10/01. PubMed PMID: 8863437.
104. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, et al. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. DNA Res. 1999;6(2):83-101, 45-52. Epub 1999/06/26. PubMed PMID: 10382966.
105. Milek I, Cigic B, Skrt M, Kaletunc G, Ulrih NP. Optimization of growth for the hyperthermophilic archaeon Aeropyrum pernix on a small-batch scale. Can J Microbiol. 2005;51(9):805-9. Epub 2006/01/05. doi: w05-060 [pii]

10.1139/w05-060. PubMed PMID: 16391661.
106. Fukuhara H, Kifusa M, Watanabe M, Terada A, Honda T, Numata T, et al. A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3. Biochem Biophys Res Commun. 2006;343(3):956-64. Epub 2006/04/01. doi: S0006-291X(06)00489-X [pii]

10.1016/j.bbrc.2006.02.192. PubMed PMID: 16574071.
107. Methe BA, Nelson KE, Deming JW, Momen B, Melamud E, Zhang X, et al. The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. Proc Natl Acad Sci U S A. 2005;102(31):10913-8. Epub 2005/07/27. doi: 0504766102 [pii]

10.1073/pnas.0504766102. PubMed PMID: 16043709; PubMed Central PMCID: PMC1180510.

108. Otgonbayar GE, Eom HJ, Kim BS, Ko JH, Han NS. Mannitol production by Leuconostoc citreum KACC 91348P isolated from Kimchi. J Microbiol Biotechnol. 21(9):968-71. Epub 2011/09/29. doi: JMB021-09-11 [pii]. PubMed PMID: 21952374.
109. Tantos A, Friedrich P, Tompa P. Cold stability of intrinsically disordered proteins. FEBS Lett. 2009;583(2):465-9. Epub 2009/01/06. doi: S0014-5793(08)01040-5 [pii]

10.1016/j.febslet.2008.12.054. PubMed PMID: 19121309.
110. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, et al. Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev. 2001;65(1):44-79. Epub 2001/03/10. doi: 10.1128/MMBR.65.1.44-79.2001. PubMed PMID: 11238985; PubMed Central PMCID: PMC99018.
111. Cox MM, Battista JR. Deinococcus radiodurans - the consummate survivor. Nat Rev Microbiol. 2005;3(11):882-92. Epub 2005/11/02. doi: nrmicro1264 [pii]

10.1038/nrmicro1264. PubMed PMID: 16261171.
112. Bakermans C. Microbial Evolution under Extreme Conditions2015 10.03.2015. 219 p.
113. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, et al. Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res. 2000;28(21):4317-31. Epub 2000/11/01. PubMed PMID: 11058132; PubMed Central PMCID: PMC113120.
114. Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD, et al. Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. Nat Biotechnol. 2002;20(11):1118-23. Epub 2002/10/09. doi: 10.1038/nbt749

nbt749 [pii]. PubMed PMID: 12368813.
115. Tsang PH, Li G, Brun YV, Freund LB, Tang JX. Adhesion of single bacterial cells in the micronewton range. Proc Natl Acad Sci U S A. 2006;103(15):5764-8. Epub 2006/04/06. doi: 10.1073/pnas.0601705103. PubMed PMID: 16585522; PubMed Central PMCID: PMC1458647.
116. Hopkin M. Bacterium makes nature's strongest glue. Nature. 2006. doi: doi:10.1038.
117. Christie-Oleza JA, Miotello G, Armengaud J. High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics. 13:73. Epub 2012/02/18. doi: 1471-2164-13-73 [pii]

10.1186/1471-2164-13-73. PubMed PMID: 22336032; PubMed Central PMCID: PMC3305630.
118. Yi H, Lim YW, Chun J. Taxonomic evaluation of the genera Ruegeria and Silicibacter: a proposal to transfer the genus Silicibacter Petursdottir and Kristjansson 1999 to the genus Ruegeria Uchino et al. 1999. IJSEM. 2007;57(4):815-9. doi: doi: 10.1099/ijs.0.64568-0.
119. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12(2):85-94. Epub 1999/04/09. PubMed PMID: 10195279.
120. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A. 1984;81(1):140-4. Epub 1984/01/01. PubMed PMID: 6582470; PubMed Central PMCID: PMC344626.

121. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105-32. Epub 1982/05/05. doi: 0022-2836(82)90515-0 [pii]. PubMed PMID: 7108955.
122. Janin J. Surface and inside volumes in globular proteins. Nature. 1979;277(5696):491-2. Epub 1979/02/08. PubMed PMID: 763335.
123. Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol. 2013;30(12):2645-53. doi: 10.1093/molbev/mst157. PubMed PMID: 24037790.
124. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. Journal of proteome research. 2006;5(4):879-87. doi: 10.1021/pr060048x. PubMed PMID: 16602695; PubMed Central PMCID: PMC2543136.
125. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. Journal of proteome research. 2006;5(4):888-98. doi: 10.1021/pr060049p. PubMed PMID: 16602696; PubMed Central PMCID: PMC2533134.
126. Denning DP, Rexach MF. Rapid evolution exposes the boundaries of domain structure and function in natively unfolded FG nucleoporins. Mol Cell Proteomics. 2007;6(2):272-82. doi: 10.1074/mcp.M600309-MCP200. PubMed PMID: 17079785.
127. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. Journal of molecular evolution. 2002;55(1):104-10. doi: 10.1007/s00239-001-2309-6. PubMed PMID: 12165847.
128. Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of molecular evolution. 2002;55(3):260-4. doi: 10.1007/s00239-002-2323-3. PubMed PMID: 12187379.
129. Pavlovic-Lazetic GM, Mitic NS, Kovacevic JJ, Obradovic Z, Malkov SN, Beljanski MV. Bioinformatics analysis of disordered proteins in prokaryotes. BMC bioinformatics. 2011;12:66. doi: 10.1186/1471-2105-12-66. PubMed PMID: 21366926; PubMed Central PMCID: PMC3062596.
130. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. Journal of molecular graphics & modelling. 2001;19(1):26-59. PubMed PMID: 11381529.
131. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993;232(2):584-99. doi: 10.1006/jmbi.1993.1413. PubMed PMID: 8345525.
132. Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, et al. Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science. 2014. doi: 10.1126/science.1249252. PubMed PMID: 24674868.
133. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2004;18(11):1169-75. doi: 10.1096/fj.04-1584rev. PubMed PMID: 15284216.
134. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000;25:25-9.

135. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One. 2013;8(11):e79217. Epub 2013/11/22. doi: 10.1371/journal.pone.0079217

PONE-D-13-31980 [pii]. PubMed PMID: 24260172; PubMed Central PMCID: PMC3829842.
136. Boles E, Hollenberg CP. The molecular genetics of hexose transport in yeasts. FEMS Microbiol Rev. 1997;21(1):85-111. Epub 1997/08/01. doi: S0168-6445(97)00052-1 [pii]. PubMed PMID: 9299703.
137. Ozcan S, Dover J, Johnston M. Glucose sensing and signaling by two glucose receptors in the yeast Saccharomyces cerevisiae. EMBO J. 1998;17(9):2566-73. Epub 1998/06/20. doi: 10.1093/emboj/17.9.2566. PubMed PMID: 9564039; PubMed Central PMCID: PMC1170598.
138. Brown CJ, Todd KM, Rosenzweig RF. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol Biol Evol. 1998;15(8):931-42. Epub 1998/08/27. PubMed PMID: 9718721.
139. Tannenbaum E. A comparison of sexual and asexual replication strategies in a simplified model based on the yeast life cycle. Theory Biosci. 2008;127(4):323-33. Epub 2008/08/22. doi: 10.1007/s12064-008-0049-5. PubMed PMID: 18716819.
140. Zeyl C, Curtin C, Karnap K, Beauchamp E. Antagonism between sexual and natural selection in experimental populations of Saccharomyces cerevisiae. Evolution. 2005;59(10):2109-15. Epub 2006/01/13. PubMed PMID: 16405156.
141. Field CM, Kellogg D. Septins: cytoskeletal polymers or signalling GTPases? Trends Cell Biol. 1999;9(10):387-94. Epub 1999/09/11. doi: S0962-8924(99)01632-3 [pii]. PubMed PMID: 10481176.
142. Madden K, Snyder M. Cell polarity and morphogenesis in budding yeast. Annu Rev Microbiol. 1998;52:687-744. Epub 1999/01/19. doi: 10.1146/annurev.micro.52.1.687. PubMed PMID: 9891811.
143. Carroll CW, Altman R, Schieltz D, Yates JR, Kellogg D. The septins are required for the mitosis-specific activation of the Gin4 kinase. J Cell Biol. 1998;143(3):709-17. Epub 1998/11/13. PubMed PMID: 9813092; PubMed Central PMCID: PMC2148151.
144. Barral Y, Parra M, Bidlingmaier S, Snyder M. Nim1-related kinases coordinate cell cycle progression with the organization of the peripheral cytoskeleton in yeast. Genes Dev. 1999;13(2):176-87. Epub 1999/01/30. PubMed PMID: 9925642; PubMed Central PMCID: PMC316392.
145. Park HO, Sanson A, Herskowitz I. Localization of Bud2p, a GTPase-activating protein necessary for programming cell polarity in yeast to the presumptive bud site. Genes Dev. 1999;13(15):1912-7. Epub 1999/08/13. PubMed PMID: 10444589; PubMed Central PMCID: PMC316924.
146. de Oliveira IM, Henriques JA, Bonatto D. In silico identification of a new group of specific bacterial and fungal nitroreductases-like proteins. Biochem Biophys Res Commun. 2007;355(4):919-25. Epub 2007/03/03. doi: S0006-291X(07)00326-9 [pii]

10.1016/j.bbrc.2007.02.049. PubMed PMID: 17331467.
147. Forestier C, Frangne N, Eggmann T, Klein M. Differential sensitivity of plant and yeast MRP (ABCC)-mediated organic anion transport processes towards sulfonylureas. FEBS Lett. 2003;554(1-2):23-9. Epub 2003/11/05. doi: S0014579303010640 [pii]. PubMed PMID: 14596908.

148. Goldberg T. HM, Hamp T., Karl T., Yachdav G., Ahmed N., Altermann U., Angerer P., Ansorge S., Balasz K. LocTree3 prediction of localization. Nucleic Acids Res. 2014. doi: 10.1093/nar/gku396.
149. Ramilowski JA, Goldberg T, Harshbarger J, Kloppman E, Lizio M, Satagopam VP, et al. A draft network of ligand-receptor-mediated multicellular signalling in human. Nature communications. 2015;6:7866. Epub 2015/07/23. doi: 10.1038/ncomms8866. PubMed PMID: 26198319.
150. Vicedo E, Schlessinger A, Rost B. Environmental pressure may change the composition protein disorder in prokaryotes. PloS one. 2014. Epub 2015/07/03. doi: 10.1371/journal.pone.0133990.

Acknowledgements

This dissertation has been designed; elaborate und worked during 5 years of research and several interesting experiences. I wish to use this opportunity to thanks all the people who made it possible.

First and foremost I want to thank Burkhard Rost for his support in many ways. I'm very grateful to him for accepting me into his group and for his motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Also, I'm in debt to my fellows colleagues Avner Schlessinger, Markus Schmidberger, Marco Punta, László Kaján, Christian Schäfer Arthur Dong, Miguel Cejuela, Tobias Hamp and Tatyana Goldberg for their stimulating discussions, collaborations and for having shared their knowledge with me and special thanks for Edda Kloppmann and Shaila Rössle-Blank for their helpfully comments, suggestions and corrections and most important their friendship.

Special thanks to Marlena Drabik and Inga Weise for the administrative part and Timothy Karl for the technical support, stimulating conversations and snack breaks rendered with jazz music.

Thanks also to the anonymous reviewers and to all those who deposit their experimental data in public databases, and to those who maintain these databases, without them this thesis could not have been realized.

My deepest heartfelt appreciation goes to all my family around the world and all the wonderful people that I have the opportunity to meet during this life but particularly I would like to thank my parents Andres y Maria, my brother Andres and his wife Elisa and my wonderful nice Andrea and nephew Armando for their endless physical and moral support and for their patience throughout all my hectic life but particularly by working on this thesis.

Appendix

The manuscripts of the following peer-reviewed publications have been appended:

- **Esmeralda Vicedo**, Avner Schlessinger & Burkhard Rost: **Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes**. PLoS One 2015, 10:e0133990.

- **Esmeralda Vicedo**, Zofia Gasik, Yu-An Dong, Tatyana Goldberg & Burkhard Rost: **Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock**

Summaries of the publications and my individual contributions are as follow

Environmental pressure may change the composition protein disorder in prokaryotes

The work presents an analysis of regions of long disorder in proteins (also referred to as natively unstructured or intrinsically disordered). It is well known that such regions are substantially more abundant in more complex organisms than in simpler ones, e.g. much more in eukaryotes than in prokaryotes. Here, we focus on disorder in prokaryotes.

Many prokaryotes have adapted to incredibly extreme habitats. How do such extremophiles differ in their genome with respect to their non-extremophile relatives? Here, we reveal that differences between organisms from distinct habitats are imprinted upon a single feature of protein structure, namely the fraction of proteins with long regions that are predicted to be disordered. In particular, we use different approaches and methods to analyze disorder abundance in whole genomes representing organisms from diverse habitats and found a correlation between protein disorder and the extremity of the environment. We conclude that the extremity of the organism environment has a considerable influence over the total content of intrinsically disordered proteins in the studied organisms, more than phylogeny.

The study was conceived and designed by myself, Avner Schlessinger and Burkhard Rost. I carried out necessary background research. The programming was performed by myself with help of Avner Schlessinger. All calculations were done by myself with help of Burkhard Rost. The resulting data was analyzed by myself, Avner Schlessinger and Burkhard Rost. The manuscript was drafted by myself, Avner Schlessinger and Burkhard Rost. The pictures were provided by myself and Burkhard Rost.

Protein disorder reduced in Saccharomyces cerevisiae to survive heat shock

Recent experiments established that a culture of Saccharomyces cerevisiae (baker's yeast) survives sudden heat shock by specifically duplicating chromosome III and two fragments of chromosomes IV and XII. This solution is "temporary" in the sense that is realized "only" for the first 400-1200 generations of the surviving cells; after some 1200 generations more detailed solutions are found.

Our manuscript establishes that heat shock proteins (HSPs) are not significantly abundant in the duplicated regions, hence, cannot be the answer for why these particular regions have been duplicated. Instead, we hypothesized that the reduction of proteins with long regions of disorder might help to acquire heat resistance. Indeed, the duplication was substantially depleted of disorder. In this view, the reduction of disorder could be perceived as some sort of "buffer". We analyzed candidates for processes that are specifically over-represented in the duplication and that could explain the advantage for survival. We identified several interesting candidates, but could not draw a convincing hypothesis.

The study was conceived and designed by myself, Zofia Gasik and Burkhard Rost. I carried out necessary background research. All calculations were done by Zofia Gasik under my supervision, Yu-An Dong and Tatyana Goldberg. The resulting data was analyzed by myself, Zofia Gasik, Tatyana Goldberg and Burkhard Rost. The manuscript was drafted by myself, Zofia Gasik, Yu-An Dong, Tatyana Goldberg and Burkhard Rost.

PLOS | ONE

RESEARCH ARTICLE

Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes

Esmeralda Vicedo[1,2]*, Avner Schlessinger[3], Burkhard Rost[1,4,5]

1 TUM, Department of Informatics, Bioinformatics & Computational Biology—i12, Boltzmannstr. 3, 85748 Garching, Munich, Germany, **2** TUM Graduate School of Information Science in Health (GSISH), Boltzmannstr. 11, 85748 Garching, Munich, Germany, **3** Icahn School of Medicine at Mount Sinai, Department of Pharmacology and Systems Therapeutics, One Gustave L. Levy Place, Box 1603, New York, New York, 10029, United States of America, **4** Institute of Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching, Munich, Germany, **5** Institute for Food and Plant Sciences WZW Weihenstephan, Alte Akademie 8, Freising, Germany

* assistant@rostlab.org

OPEN ACCESS

Citation: Vicedo E, Schlessinger A, Rost B (2015) Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes. PLoS ONE 10(8): e0133990. doi:10.1371/journal.pone.0133990

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: September 11, 2014

Accepted: July 3, 2015

Published: August 7, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Many prokaryotic organisms have adapted to incredibly extreme habitats. The genomes of such extremophiles differ from their non-extremophile relatives. For example, some proteins in thermophiles sustain high temperatures by being more compact than homologs in non-extremophiles. Conversely, some proteins have increased volumes to compensate for freezing effects in psychrophiles that survive in the cold. Here, we revealed that some differences in organisms surviving in extreme habitats correlate with a simple single feature, namely the fraction of proteins predicted to have long disordered regions. We predicted disorder with different methods for 46 completely sequenced organisms from diverse habitats and found a correlation between protein disorder and the extremity of the environment. More specifically, the overall percentage of proteins with long disordered regions tended to be more similar between organisms of similar habitats than between organisms of similar taxonomy. For example, predictions tended to detect substantially more proteins with long disordered regions in prokaryotic halophiles (survive high salt) than in their taxonomic neighbors. Another peculiar environment is that of high radiation survived, e.g. by *Deinococcus radiodurans*. The relatively high fraction of disorder predicted in this extremophile might provide a shield against mutations. Although our analysis fails to establish causation, the observed correlation between such a simplistic, coarse-grained, microscopic molecular feature (disorder content) and a macroscopic variable (habitat) remains stunning.

Introduction

Disordered regions might contribute to complexity of an organism

We refer to disordered regions as those long stretches of consecutive residues in proteins that do not adopt well-defined three-dimensional (3D) structures in isolation [1]. Proteins with long disordered regions encompass some unique biophysical characteristics which allow them

to bind to several different partners, often at different times and under different cellular conditions [2]. Typically regions with at least 30 consecutive residues predicted as disordered are considered as "long". Computational predictions have noted an overabundance of disordered regions in protein interaction hubs [3–7] and in transcriptional master regulators [8, 9]. Proteins with disordered regions appear to be particularly abundant in processes such as transcription, translation, signal transduction, and macromolecular transport through the nuclear pore complex [4, 10, 11]. All these observations support the to some degree oversimplified view of disordered regions as building blocks for system complexity [1]. On the level of kingdoms: 10–20% of all proteins from prokaryotes have at least one long disordered region, while 20–50% of all eukaryotic proteins do [1, 12, 13]. Recent comparative proteomics studies have strengthened the link between disorder and organism complexity, e.g. disordered regions in ancient branching eukaryotes appear to differ from those in other eukaryotes [14–16].

Comparative proteomics reveals new evolutionary links

How does the complexity of an organism evolve? Do humans share a minimal set of genes with bacteria and have all others evolved for non-bacteria specific functions [17]? These two questions have been pursued by many comparative genomics studies [18] for many years; the final explanation is still being sought after. One approach to comparing genomes is to focus on characteristics of proteins. For example, combining analysis of sequence, structure, expression and evolutionary relationship information of multiple protein data sets from yeast, mouse and human, evidence could be found about the relationships between divergence in the length of disordered regions and changes in the protein functions [19]. A modification of the length of disordered regions in paralog proteins might provide a simple evolutionary mechanism for protein degradation rates. As many of these affected paralogs were participating in protein signaling pathways, the cellular function and phenotype of the cells would also be influenced by these changes [20–22]. It is also a well-known fact that intertwined helices (coiled-coils) are highly over represented in eukaryotes [23]. Helices might constitute excellent evolutionary building blocks as they can form exclusively from local internal molecular interactions [24]. Through the application of prediction methods, we can integrate this useful information to compare structural features across species for entire proteomes [11, 17, 23, 25–28]. In our study we focus on the study of simple, average features from predictions that can be obtained for entire organisms.

How do prokaryotic proteins adapt to the extreme?

It appears intuitive to assume that increasing the internal inter-residue bonds in a protein raises its stability at high temperature. Several studies have, indeed, reported correlations between thermal stability and features such as a high contact density and unusual numbers of hydrogen bonds [29, 30]. A difference in the average amino acid composition was found when considering in more detail the amino acid composition, the sequence of proteins from thermophiles and those of mesophiles [31]. Protein structures from thermophiles such as *Pyrococcus horikoshi* OT3 have been reported to contain more intra-helical salt bridges than their homologues in mesophiles [32]. These salt bridges are an important factor stabilizing thermophilic proteins [30]. All these findings suggest that diverse factors determine thermostability [33]. Psychrophiles live in the extreme cold. Recent studies have suggested that proteins from psychrophiles increase their flexibility and accessibility and might thereby hinder freezing [34]. Proteins from *halobacteria* (salty habitats) also exhibit unique characteristics such as low hydrophobicity, excess of acidic residues, depletion of cysteine residues and reduced propensities for helix

formation [35]. All these observations induced us to hypothesize that protein disorder might somehow correlate with habitat.

Assuming that protein disorder plays a marginal role in prokaryotes, most studies have focused on eukaryotes. Here, we zoomed into protein disorder abundance across prokaryotes. Specifically, our first question was whether the overall percentage of proteins with long regions of protein disorder is associated with organism habitat, or alternatively, with taxonomic distance. Put differently: are two proteomes more similar in their disorder content when they are related by evolution or when they live in similar habitats? We predicted disorder through several *in silico* methods applied to about 46 organisms that thrive in different habitats. Overall, we claim to have established a stronger correlation between disorder and habitat than between disorder and taxonomy for the same control set. Furthermore, our results appeared more compatible with the idea of "gradual adaptation" than with that of "gradual leap", i.e. disorder regions were added to many proteins, rather than introducing a few new, organism-specific proteins with disordered regions.

Methods

Data

The UniProt database [36] provided the complete proteome sequence data at the basis of our study. We removed all duplicates (giving priority to longer proteins) and applied no other filtering. Our analysis considered 46 organisms with a total of 225,550 proteins (S1 Table). The organisms sampled the most extreme habitats and their closest completely sequenced relatives. We also included a few selected eukaryotes for comparison.

Most information used to classify organisms was taken from GOLD (Genomes Online database version 2011-09-23 [37]). We avoided pathogens, parasites, and other biotic relationships to build a "simplified" subset of organisms. We classified into the following types of environment (S1 Table) [38–40]: thermophiles (optimal growth at 45–80° C), hyperthermophiles (optima >80° C), pychrophiles (optimal growth at about 15° C, a maximal temperature for growth at about 20° C, and a minimal temperature for growth at 0° C or below), psychrotolerants (organisms that are not considered as pyschrophile but have the capability for growth at 0° C or close to 0° C), halophiles (optimal growth in salt solutions, i.e. from 25% NaCl up to saturation), alkaliphiles (optimal growth around pH>8), mesophiles (including bacteria and archaea from "normal" environments). Eukaryotes were considered as a different group as they have a different content of disorder [1].

Disorder prediction

We used prediction methods that were developed based on different concepts and capture different "flavors" of protein disorder [6, 41, 42]. Therefore when analyzing the predicted amount of disordered proteins in an organism, it is possible to obtain distinct values depending on the predictor. IUPred uses pairwise statistical potentials of residue contacts [43, 44] and has been presented as an unbiased and robust predictor even for organisms living in extreme habitats [45, 46]; Meta-Disorder (MD) [42] and NORSnet [6] are neural network-based methods that use evolutionary information and other predicted features. MD combines several original prediction methods including NORSnet, with evolutionary profiles and sequence features that correlate with protein disorder such as predicted solvent accessibility and protein flexibility. NORSnet is focused on the identification of long disordered loops (no regular secondary structure, namely "loopy disorders"); it is optimized without using any experimental data on disorder. Disordered regions that are not predicted to be "loopy" are considered as "regular" disordered regions.

There are many alternatives how to compile overall averages for protein. We analyzed almost the entire resulting data avalanche and found most alternatives to be redundant. Therefore, we focused on as few alternatives as possible; we included different views only if they provided important additional information. In particular, we considered three thresholds to define "long disorder": **%long30**, is the percentage of proteins with at least one region of ≥30 consecutive residues predicted as disordered (**%long50** and **%long80** were the same with length thresholds at ≥50 and ≥80, respectively). We also investigated another extreme concept, in particular that of a protein that is **completely disordered** (S1 Fig): if a protein had no single region that we could perceive as a "nucleation site" for adopting regular structure, we considered this protein as completely disordered. Operationally, we first removed any prediction of disorder that spanned over fewer than five residues; next we searched any region without predicted disorder over 30 consecutive residues. If we found no such region, and if we also found at least one region with ≥30 consecutive residues predicted as disordered, we considered the protein to be completely disordered. All thresholds were tested with three prediction methods, concretely MD, NORSnet and IUPred. To simplify comparisons between these three, we replaced their raw scores by Z-scores, i.e. gave the score as a deviation from the average in units of one standard deviation:

$$z_{(o,\,M)} = \frac{\mathrm{raw}_{(o,\,M)} - \langle \mathrm{raw} \rangle_{(\text{all organisms},\,M)}}{\sigma_{(\text{all organisms},\,M)}} \qquad \text{(Eq1)}$$

where $z_{(o,M)}$ is the Z-score for a particular method M and organism o, $\mathrm{raw}_{(o,M)}$ is the raw score of prediction method M for organism *o* (e.g. the percentage of proteins with at least one region of long disorder in *o*), $\langle \mathrm{raw} \rangle_{(\mathrm{allorganisms},M)}$ is the average over the raw scores for method M over all organisms, and $\sigma_{(\mathrm{allorganisms},M)}$ is the standard deviation for the distribution of the raw scores predicted for all organisms by method M. Positive Z-scores imply a disorder content higher than the mean, negative scores lower than the mean. We compiled averages and standard deviations over a set of 1,613 complete prokaryotic proteomes from UniProt (with almost 90% of the sequences predicted by the three predictors) in order to have a Z-score calculated independently of the samples selected and to give more information compared to the total of the 1,613 organisms. Eukaryotes were not included in this computation due to the difference in disorder content [13]; they were considered separately for the analysis. The calculated means (*ave*) and standard deviations (*sd*) for **"%long30"** were: MD_{ave} = 14.6%, MD_{sd} = 4.2%; $\mathrm{NORsnet}_{ave}$ = 2.5%, $\mathrm{NORsnet}_{sd}$ = 2.0%; IUPred_{ave} = 7.5% and IUPred_{sd} = 5.5% (for other approaches see S3–S5 Tables).

Tree of life

We constructed and visualized the tree of life using the interactive *Tree of Life* (ITOL) webserver [47, 48]. Taxonomic identifiers for the organisms were taken from UniProt and uploaded into the NCBI taxonomy browser [49, 50] to automatically generate a phylogenetic tree in *phylip* format [51]. The resulting tree was visualized using the "Multi-value Bar Chart" a circular mode of ITOL.

Defining homology

In order to identify phylogenetic relations such as the homology of proteins between the thermophile *Pyrococcus horikoshii OT3* [52] and the model organism for the study of life in permanently cold environments *Colwellia psychrerythraea 34H* [53], we applied the following *ad hoc* procedure: We blasted [54] all protein sequences from one organism against all from the other. For each resulting alignment we calculated the HSSP-value (HVAL) [55–57], which measures

sequence similarity by combining alignment length and percentage of pairwise sequence identity. For instance, HVAL = 0 corresponds to about 22% pairwise sequence identity for alignments over 250 residues. As a result of our procedure, proteins can have multiple homologues. Due to technical concerns, we grouped all relations found avoiding the problem in the distinction between paralogs and orthologs [58, 59].

Statistical tests

In addition to the similarity between proteins from two organisms, we also assessed the statistical significance of disorder content comparisons between organisms with similar habitat (S1 Table) and with similar phylogeny (S14 Table). In particular, we applied the Kruskal-Wallis test (H-test) [60, 61], the Wilcoxon signed-rank test [62–65] and the Brown–Forsythe Levene's test (also known as Levene's test) [66, 67] (S2 Fig). The non-parametric Kruskal-Wallis test compares the shape of the distributions between two or more unmatched groups for nominal variables of small and unequal sample size and determines whether the distributions of the groups are identical (null hypothesis) [60, 62, 63]. The pairwise Wilcoxon signed-rank test is a nonparametric test for matched or paired data to assess whether the differences of the median between pairs of observations is zero [62–65]. The Levene's test is a non-parametric test that also works for non-Normal (non-Gaussian) distributions; it determines if all variances between groups are zero (null hypothesis, $\alpha = 0.05$) [66, 67]. For all the statistical tests, we used the median for each group either habitat or phyla, calculated from the protein disorder content of the organisms belonging to this group.

The Kruskal-Wallis test does not assume a normal distribution for the data but homoscedasticity (not significant differences between the group variances) [60, 61] therefore first, we performed the Levene's test of equality of variances (S2 Fig). If the Levene's test failed for the overall comparison across the groups, then we performed pairwise comparisons between the groups (S2 Fig). For those groups for which the null hypothesis (equal variances) is accepted, a pairwise Wilcoxon signed-rank test will be applied as alternative to the Kruskal-Wallis Test (null hypothesis: groups have equal distribution; $\alpha = 0.05$; S2 Fig). The groups rejecting the null hypothesis and therefore presenting a significant difference of disorder content distribution were all marked with asterisks ($P < 0.05$ with * and $P < 0.005$ with **). The pairwise Wilcoxon signed-rank test was also applied when the Kruskal-Wallis test failed for the overall comparison test (accept alternative hypothesis, i.e. at least one group in the population for which the distribution of disordered protein contents differs from the others) and after the null hypothesis of the pairwise homogeneity Levene's test was accepted (S2 Fig). Furthermore, habitat is a complex reality defined by a variety of ambient conditions and organism properties which have to be studied separately. For that we also analyzed, some of the general properties of the organisms (metadata) included by the GOLD database [37]. For the statistical analysis groups containing less than two samples were not considered. All analyses were performed using the R software (statistical packages car and stats) [66, 68].

Results & Discussion

Salty habitats are dominated by high disorder

Halophiles thrive in salt-saturated habitats. The percentages of proteins predicted with long disorder in the two halophilic archaea *Halobacterium sp. NRC-1* [69] and *Haloarcula marismortui ATCC 43049* [70] both reached levels around 20–28% (percentage of proteins with at least one region with >30 consecutive residues predicted to be disordered by MD and IUPred). This was much higher than average (Z-scores Fig 1A, note Z-score = 0 implies 'like average', +1/-1: imply values one standard deviation above/below average) and much higher than the

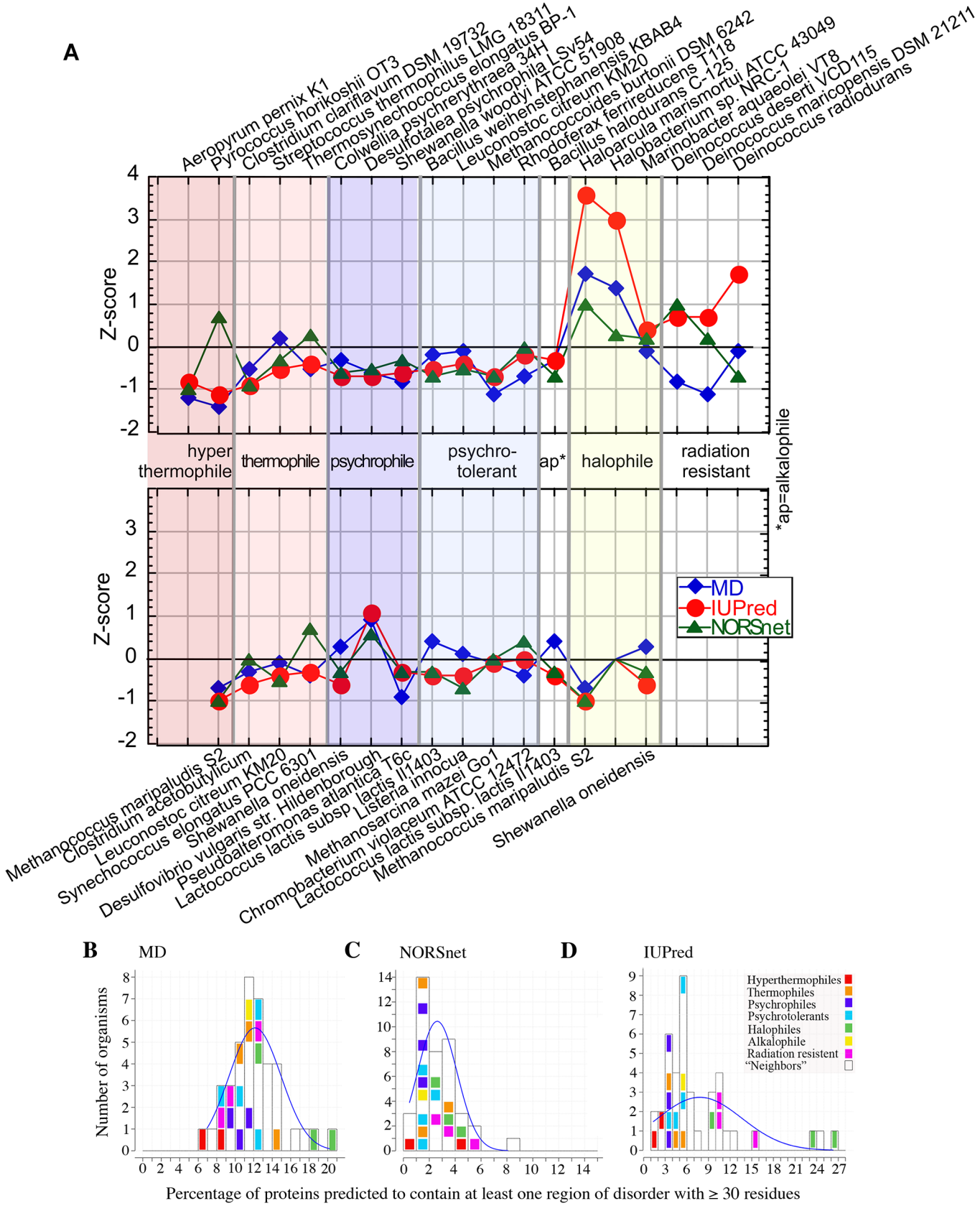
A
Aeropyrum pernix K1
Pyrococcus horikoshii OT3
Clostridium clariflavum DSM 19732
Streptococcus thermophilus LMG 18311
Thermosynechococcus elongatus BP-1
Colwellia psychrerythraea 34H
Desulfotalea psychrophila LSv54
Shewanella woodyi ATCC 51908
Bacillus weihenstephanensis KBAB4
Leuconostoc citreum KM20
Methanococcoides burtonii DSM 6242
Rhodoferax ferrireducens T118
Bacillus halodurans C-125
Haloarcula marismortui ATCC 43049
Halobacterium sp. NRC-1
Marinobacter aquaeolei VT8
Deinococcus deserti VCD115
Deinococcus maricopensis DSM 21211
Deinococcus radiodurans
Z-score
hyper thermophile
thermophile
psychrophile
psychro-tolerant
ap*
halophile
radiation resistant
*ap=alkalophile
MD
IUPred
NORSnet
Methanococcus maripaludis S2
Clostridium acetobutylicum
Leuconostoc citreum KM20
Synechococcus elongatus PCC 6301
Shewanella oneidensis
Desulfovibrio vulgaris str. Hildenborough
Pseudoalteromonas atlantica T6c
Lactococcus lactis subsp. lactis Il1403
Listeria innocua
Methanosarcina mazei Go1
Chromobacterium violaceum ATCC 12472
Lactococcus lactis subsp. lactis Il1403
Methanococcus maripaludis S2
Shewanella oneidensis
B
MD
C
NORSnet
D
IUPred
Number of organisms
Hyperthermophiles
Thermophiles
Psychrophiles
Psychrotolerants
Halophiles
Alkalophile
Radiation resistent
"Neighbors"
Percentage of proteins predicted to contain at least one region of disorder with ≥ 30 residues

Fig 1. Distribution of disorder content in different organisms. Fractions of proteins with long regions of disorder (here ≥30 consecutive residues) were predicted by three prediction methods (MD, NORSnet and IUPred). **(A)** The raw values are standardized using the Z-scores (Eq 1; mean and standard deviation σ from a 1613 prokaryotes calculated for each method; positive: higher than the mean; negative: below the mean; integers +/- N imply N*σ above/below the mean). The top panel shows the extremophiles; the lower panel shows the closest phylogenetic relative for each extremophile in the top panel (for relatives discussed in the text and left out for clarity from the figure, for all studied organisms S3 Fig). The archaeal halophiles *Haloarcula marismortui ATCC 43049* and *Halobacterium sp. NRC-1* were predicted with the highest content of proteins with long disorder. Conversely, the archaeal thermophile *Aeropyrum pernix K1* was one of the organisms predicted with the lowest disorder. The taxonomic neighbors section compares the disorder predicted for the closest relatives of the extremophiles. **(B-D)** Mapping of disorder protein content predictions for all organisms for each prediction method (B: MD [42], C: NORSnet [6], and D: IUPred Clearly, all three methods put the thermophiles on the left (less disorder), while the halophiles appear on the right (high disorder). The blue curves are Gaussian fits based on the mean and σ of our data.

doi:10.1371/journal.pone.0133990.g001

values for their closest taxonomic relative *Methanococcus maripaludis* S2 [71] (Z-scores<-0.5 Fig 1A) that does not survive in high salt. The same tendency was observed for the other methods and thresholds (S7 and S8 Tables).

The difference in disorder abundance between the halophilic bacterium *Marinobacter aquaeolei VT8* [1] (Z-score around 0, Fig 1A) and its taxonomic relative *Pseudoalteromonas atlantica T6c* [72] (Z-score around -0.5, Fig 1A) was not as pronounced as for the archaea, but it confirmed the "high disorder in salt" trend for bacteria. The difference in disorder between halophile and relative was slightly higher for longer disorder (S8 Table and S4 and S5 Figs). When considering the percentage of proteins considered as completely disordered (S1 Fig), the difference increased (S2 Table vs. S9 Table). The difference was the same in relative terms for a method that detects only long loops (no regular secondary structure, such as NORSnet) as disorder, although the content for that method dropped significantly (NORSnet in Fig 1A). These observations across different phyla might suggest the increase in disordered regions as one means for prokaryotes to cope with high salt-conditions. This result has been reported before [45, 73]. New here is the relation between phylogeny (closest relatives) and extremity of habitat (high salt).

Is disorder slightly lower in hot habitats?

Organisms surviving in extreme heat have been reported to have rather low levels of disorder content before ([45]). The group of Peter Tompa–[45]—also reported a low content of disorder in organisms surviving the cold and put these results into perspective of evolutionary relatives. Here, we repeated their analysis in a slightly wider context, largely confirming their findings.

The hyperthermophile *Pyroccocus* [74] might be the most studied organism living in very high temperature (close to 100°C) and greater sea depth than other archaea (pressures reaching 200 bar, i.e. ~200 times what we live in). At least for two of the methods we analyzed, *Pyroccocus horikoshii OT3* [75] was predicted with very little long disorder (>30 residues, Fig 1A: <-1, i.e. over one standard deviation below average). The closest relative, *Methanococcus maripaludis S2*, was predicted with similar low disorder (Z-score around -1 Fig 1A). The optimal growth temperature for *Methanococcus maripaludis* is 35–40°C, i.e. "normal", and it is isolated from salt marsh sediments. Following our simple logic, we expect two reasons for *Methanococcus* to have higher disorder than *Pyroccocus*: salt (higher disorder) and less heat (higher disorder). For our method predicting loopy disorder, the trend was even inversed. We failed to explain why we did not observe this.

Aeropyrum pernix K1 (isolated from sulfur-rich under-sea vents in Japan) [76–78] is another hyperthermophile archeae. Like *Pyroccocus*, *Aeropyrum* was predicted with very little disorder (Z-score ~-1, Fig 1A). This was similar to other hyperthermophiles that we sampled. Analogous to the halophiles, the "loopy" disorder predicted by NORSnet, was even lower for these hyperthermophiles than the "regular" disorder. While we might jump into suspecting that shortening connections between regular secondary structure segments (helices and

strands) might protect against heat and high salt, we should speculate with care because this seems incompatible with the prediction of "loopy disorder" for *Pyroccocus* (Fig 1A).

Disorder seems not higher in cold habitats

Colwellia psychrerythraea 34H [53] is considered as an obligate psychrophile marine bacterium, i.e. it needs very low temperatures (-1°C to +10°C) to grow; it can support high pressures in the deep sea. Its predicted disorder was below average (Z-score about -0.5, Fig 1A). *Leuconostoc citreum KM20* [79] is considered to be a psychrotolerant antimicrobial producer (used for fermentation of kimchi). It grows optimally at 30°C, but can also be cultivated at significantly higher temperatures. Its predicted disorder was also below average (Z-score about -0.5; Fig 1).

A recent study provided experimental evidence that proteins with long disordered regions can be more stable in cold temperatures than globular proteins [80]. Our predictions for entire genomes seemed incompatible with the concept that such a solution would be imprinted upon the entire proteome. If anything, our analysis of psychrophiles confirmed previous findings that organisms in cold habitats have less disorder than average ([45]).

Is high disorder protecting from radiation?

Deinoccocus radiodurans R1 [81, 82] is often jokingly referred to as "Conan the bacterium" because it tolerates many extreme conditions including radiation, cold, dehydration, heat and high acidity. We predicted a high abundance of protein disorder in this bacterium (Z-score between 0 and 2: Fig 1A). We only found two taxonomic neighbors of *Deinoccocus radiodurans*: *Deinococcus deserti* and *Deinoccus maricopensis*. Both also sustain high radiation and live in the dry: *Deinococcus deserti* and *Deinoccus maricopensis* (Z-scores >0 for IUPred, Fig 1A). The 'high radiation' habitat was particularly inconsistent between the three prediction methods: e.g. MD predicted the opposite (Fig 1A). Inconsistency between prediction methods might suggest taking the correlation 'high radiation—high disorder' with a grain of salt. Conversely, we might argue for the opposite: IUPred, MD, and NORSnet rely on partially orthogonal information. This independence might imply that some reality might be discovered by only one of the methods, namely the one better able to capture that reality.

No clear trends for other disorder outliers

Finally, we analyzed the disorder abundance in prokaryotes that live in other extreme habitats including high pH (*Bacillus halodurans* [83], disorder below average, Fig 1A) and changing environments (*Shewanella oenidenses* [84], disorder around average, Fig 1A). However, so far we failed to notice significant trends (Fig 1A). Moreover, we failed to explain why some mesophiles were outliers (higher or lower content of disordered proteins). For example, *Caulobacter vibrioides* (also known as *Caulobacter crescentus*) [56] was predicted with high disorder (Z-score one standard deviation above average, Fig 1A) without any apparent reason. *Caulobacter* secretes Nature's strongest glue [85, 86]. This might point to another important role for high content of disorder. *Streptomyces coelicor* was also predicted with higher than average disorder (Z-score >1, Fig 1A); this might be explained by its complex life cycle and production of antibiotics (their products are pharmaceutically used as anti-tumors agents, immunosuppressants and antibiotics).

Ruegeria pomeroyi DSS-3 [87] (originally classified as *Silicibacter pomeroyi* [88]) was predicted with very low disorder (Z-score about -1, Fig 1A). Its taxonomic neighbor, *Rhodobacter sphaeroides 2.4.1*, was predicted at above average disorder (Z-score>0, Fig 1A). *Ruegeria* was isolated from seawater off the US-Southeast coast; it lives at 10–40°C and grows with and

without carbon monoxide (CO) as carbon source. We cannot explain the low protein disorder content predicted for *Ruegeria*.

Detailed analysis of corresponding homologues brings new insights

We calculated disorder abundance in organism specific and homologues of two model organisms representing two extreme temperature environments, using various thresholds in terms of sequence similarity to define homology (Table 1). The aim was to analyze whether the aligned region of the corresponding homologues from two opposing extremophiles (heat/cold) includes the disordered region or not. In particular, we compared the homologues between the low-temperature/low-disorder psychrophile *Colwellia psychrerythraea 34H* and the high-temperature hyperthermophile *Pyrococcus horikoshii OT3*.

At pairwise protein similarity levels of HVAL≥10 (corresponding to about 30% pairwise sequence identity over 250 aligned residues), seven of the homologs with disorder in *Colwellia* (cold) had no disorder in *Pyrococcus* (heat; S9 Table); the number for the flipside control was: one protein with disorder in *Pyrococcus* and not in *Colwellia*.

Several studies investigating the effect of temperature on enzymes–which are disorder depleted as a class of proteins—showed that proteins from extremophiles (both cold and hot) adopt similar structures as their mesophilic orthologs, but use different amino acids to compensate for temperature effects [30, 31, 34]. Our analysis confirms this trend (S6A Fig), the particular choice of amino acids in whole proteomes of hyperthermophile (S6A Fig: red) and thermophile (S6A Fig: blue) were slightly different compared to that for psychrophile (S6A Fig: green) and psychrotolerant (S6A Fig: purple) organisms. However, the differences were significant at best for some particular amino acids. The strongest signal was for negatively charged amino acids such as glutamic acid (E, S6A Fig), that occurred more in heat than in cold. The

Table 1. Protein disorder overlap between related proteins in opposite extremophiles.

HVAL [a]	*Colwellia psychrerythraea* 34H (freeze)		*Pyrococcus horikoshii* OT3 (heat)	
	related [b]	*related+disordered* [c]	*related* [b]	*related+disordered* [c]
-20	75.5 ±0.2	9.5 ± 0.1	66.9 ± 0.1	5.53 ± 0.06
-10	56.4 ±0.2	6.8 ± 0.1	55.7 ± 0.2	5.04 ± 0.08
0	24.0 ±0.1	4.9 ± 0.2	30.9 ± 0.1	2.7 ± 0.1
10	5.5 ±0.1	2.6 ± 0.2	9.7 ± 0.1	0.51 ± 0.06
20	0.6 ±0.02	0.07 ± 0.02	1.28 ± 0.03	0
30	0.04±0.01	0	0.20 ± 0.01	0

a **HVAL** measures sequence similarity as the distance from the HSSP-curve [55, 89]; e.g. HVAL = 0 implies 20% pairwise sequence identity (PIDE) for >250 aligned residues [57] (or 20+N% PIDE at HVAL = N).

b **related** gives the percentage of proteins in one organism (CP: *Colwellia psychrerythraea* 34H or PH: *Pyrococcus horikoshii*) that have corresponding homologs in the other (PH or CP) at the given HVAL [a] (totals: CP = 4423 and PH = 1573). For instance, 24% of all 4423 CT proteins have a match in one of the 1573 PH proteins at HVAL≥0, while almost 31% of the PH proteins have a homolog in CP at this level of sequence relation. One standard error is marked as '±stderr'.

c **related+disordered** gives the percentage of proteins in one organism (CP or PH) that are related [b] and have at least one disordered region (>30 residues, prediction by MD; other methods and thresholds in SOM) in the other (PH or CP) at the given HVAL [a]. Overall MD predicts 12% of all *Colwellia psychrerythraea* 34H and 8% of all *Pyrococcus horikoshii* OT3 proteins to have at least one long disordered region (Table 1; cold = high disorder, heat = low). These numbers imply that the proteins shared between the two extremophiles from opposite ends of the temperature spectrum are depleted in disorder with respect to the entire proteome. For instance only 4.9% are related and disordered from the CP perspective at HVAL≥0 as opposed to 12% for all proteins. The more similar the homologs the more the related proteins were selected to not contain disorder. One standard error is marked as '±stderr'.

doi:10.1371/journal.pone.0133990.t001

situation was, however, almost inversed for the negatively charged and slightly less acidic aspartic acid (D, S6A Fig). Glutamic acid might be abundant in heat to favor electrostatic interactions in these proteins and thereby increase their stability [90]. The only other amino acid occurring more often in thermophiles and hyperthermophiles was tyrosine (Y, S6A Fig). On the other hand, the hydrophobic methionine (M, S6A Fig) was over-represented in both psychrophiles and psychrotolerants. When grouping all amino acids in two classes (hydrophobic/not) using different hydrophobicity scales (Eisenberg and Weiss [91], Kyte-Doolittle [92], and Janin [93]), we could confirm the observation [34] that psychrophiles have less hydrophobic residues than hyperthermophiles (but not less than thermophiles): the differences we observed between the antipodes (cold/heat, S6B Fig) were insignificant (Z-score between -0.05 and -0.1-for the psychrophiles vs. 0.04–0.2 for the hyperthermophiles).

Let us nevertheless assume that our findings had established the amino acid differences to be significant so that organisms could adapt to opposite temperature scales by altering the amino acid composition in all proteins. If true, the proteins that are shared between different extremophiles would be aligned to each other independently of their disordered regions. If these observations were always true, all seven disordered regions from *Colwellia* would likely fall within the aligned regions from *Pyroccocus*. The discrepancy between the expected 32 disordered proteins and the observed 7 (S12 Table) could be explained by the fact that proteins from thermophilic organisms might "tighten the loops" [30] to increase thermostability, and psychrophilic proteins might "loosen the loops", i.e. might use more flexible loops to compensate for freezing effects. This could explain the long gaps in the alignments between the two homologous proteins that far exceed those needed to align each of them to its mesophilic relative. An alternative explanation is that these unaligned, disordered regions from *Colwellia* function as antifreeze proteins, which are unique to psychrophiles, and are capable of binding ice crystals using a large surface, thereby lowering the temperature, or changing the physico-chemical surroundings of the organism [34].

Overall, it seems likely that the difference in disorder between *Colwellia* and *Pyroccocus* on opposite sides of a tremendous temperature spectrum largely originated from homologous proteins that kept their overall shape with some modifications to adapt to extreme climates. These modifications may include shorter loops, less surface area and more compact proteins in thermophiles, and exceptionally flexible proteins in psychrophiles. Our comparison between the two opposite (cold/heat) extremophiles suggested that overall the total disorder composition was affected by many small rather than by a few big changes.

Disorder differs more between habitats than between phyla

Through the application of the Kruskal-Wallis and the paired Wilcoxon-Test, we found that the habitat groups presented different distribution of disordered content for MD ($P<0.05$; S15 Table and Fig 2A) and IUPred predictions ($P<0.05$, $P<0.005$; S17 Table and Fig 2B) and for all thresholds (%long30, %long50 and %long80; Fig 2 and S7 and S8 Figs). Conversely, the phyla groups largely did not differ in any statistically significant way (Fig 2; S15–S17 Tables). Exceptions were differences in protein disorder content between the groups for NORSnet ("loopy" disorder) for all thresholds, for MD only for the middle long disordered proteins (%long50 and only for one pair of the groups in %long30; S15 Table and Fig 2A.) and for IUPred for the proteins containing long disordered regions (%long80; S15–S17 Tables and S8C Fig). Thus, the "loopy" disorder appeared more conserved than other disordered regions [94]. But why were disordered regions longer than 80 consecutives residues affected? While we lack sound explanations, we observe that other studies support the opposite [20, 95–98]. When analyzing the completely disordered proteins we found that both, phyla and habitat have an influence on the

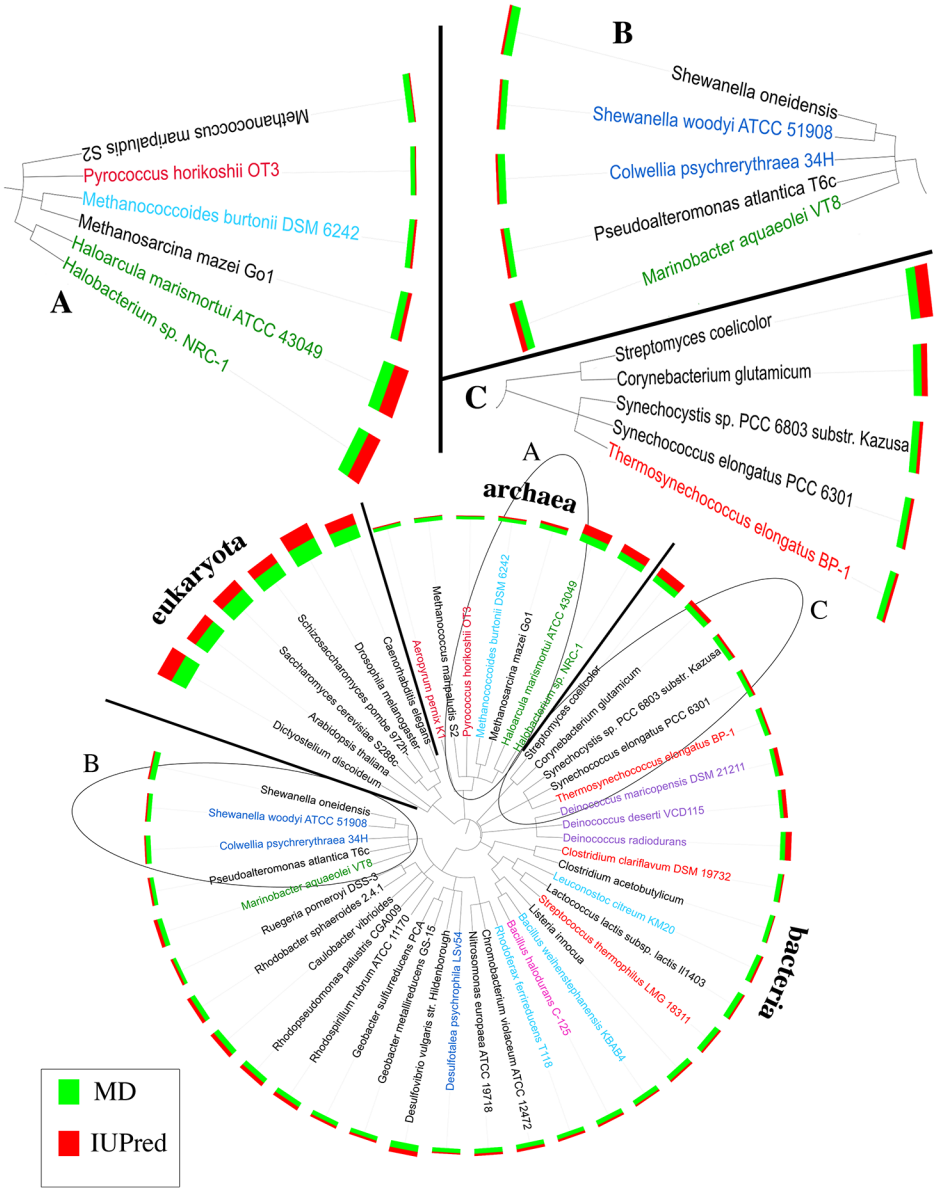

Fig 2. Protein disorder content differs for habitat, not for phyla. We represent the protein disorder content for the organisms in similar habitats (left panel) and those in the same phyla (right panel). The y-axes give the percentage of proteins with at least one region of ≥30 consecutive residues predicted as disordered by MD (A), NORSnet (B) and IUPred (C). The x-axis on the left side marks the different environmental groups (S2 Table); on the right side marks the studied phylogenetic groups (S14 Table). The groups which are significant for a paired Wilcoxon Test are marked with * (P<0.05) or ** (P<0.005).

doi:10.1371/journal.pone.0133990.g002

disorder content distribution for the IUPred and NORSnet predictions but only for disordered regions with at least 50 consecutives disordered residues (%long50 and %long80; S18 Table and S7 and S8 Figs). All those observations were confirmed when considering Z-scores (S19–S22 Tables).

The habitat is a complex reality defined by a variety of factors such as temperature, pH, energy source and metabolism (S14 Table). We tried to analyze these factors as separately as possible and in doing so we also found a significant difference in disorder content between the

organisms grouped by temperature (high temperature–low disorder; S15, S18 and S22 Tables) and by oxygen requirement (an aerobic lifestyle implied higher disorder [99, 100]; S15, S17–S19, S21 and S22 Tables). However, for the other factors (metabolism, energy source, cell shape, S15–S22 Tables) we did not observe a significant influence on disorder (content of proteins with long disordered regions). Finally, we could only suggest that in general the protein disorder abundance in proteomes is more related to environment than to phylogeny but this might be the opposite for "loopy" disorder.

Null hypothesis that disorder similar between habitats clearly rejected

Protein disorder is much more abundant in eukaryotes than in prokaryotes ([10, 101]). Nevertheless, there are substantial differences between prokaryotes (Fig 3) that appeared to correlate more between habitats than between phyla, i.e. proteins from similar habitats appeared more similar in terms of the percentage of proteins with long disordered regions than proteins with similar phylogeny (Figs 1–3). Although we reported some examples for strong correlation between habitat and disorder, we also came across many examples of organisms for which our simple hypothesis predicted the opposite of what we observed. For instance, the hyperthermophile *Pyroccocus horikoshii* was predicted with below-average disorder while its closest relative *Methanococcus maripaludis S2* was predicted with similar low disorder although it cannot survive in the heat and survives high salt which we showed to correlate with high disorder. Another conundrum originated from the detailed comparison between two organisms at opposite ends of the temperature extremity: the low-temperature/low-disorder psychrophile *Colwellia psychrerythraea 34H* and the high-temperature hyperthermophile *Pyrococcus horikoshii OT3*. The detailed comparison of corresponding related proteins ('orthologs') provided evidence that longer loops and more disorder might help to survive in the extreme cold. On the level of entire organisms we observed the opposite (and thereby confirmed previous results ([45]). May be others will bring clarity to the confusion we find in the data. While our data might not suffice to clearly prove the correlations, the data is clear enough to reject the null hypothesis (disorder not correlated between habitats). In other words, there is a signal but it might remain hidden because it might be overshadowed by other constraints for survival.

What if the signal that we report were caused by mistakes in the method? We might suspect that prediction methods have not been developed for the type of organisms for which we apply these methods here. There is little evidence for the validity of this concern. For instance, secondary structure prediction methods developed over 22 years ago ([102]) continue to correctly capture the situation for very different proteins from very different environments than had been anticipated to exist 20 years ago (disorder just being one case in point–[10]). Similarly, none of the methods that we used seems to have been optimized in any way on data specific to non-extremophiles. Another major problem coming with the diversity of disorder predictions considered for this analysis pertains to the alternative outlier or majority, i.e. should we report what one particular methods sees or should we focus on the consensus of the majority of methods. Again, there seems ample misunderstanding spread in the literature as to this matter. Some methods predicting disorder differ greatly and systematically because they capture different aspects of disorder. Differences between two data sets captured by one method and not by two others may point to the exact reason why that 'outlier' method correctly captures a reality missed by the other two. Given the heterogeneity of the phenomenon protein disorder, this seems a very likely interpretation when comparing different methods. In our example, this might indicate that the IUPred prediction that radiation resistant correlates with high disorder might be more helpful than the MD prediction of the opposite trend.

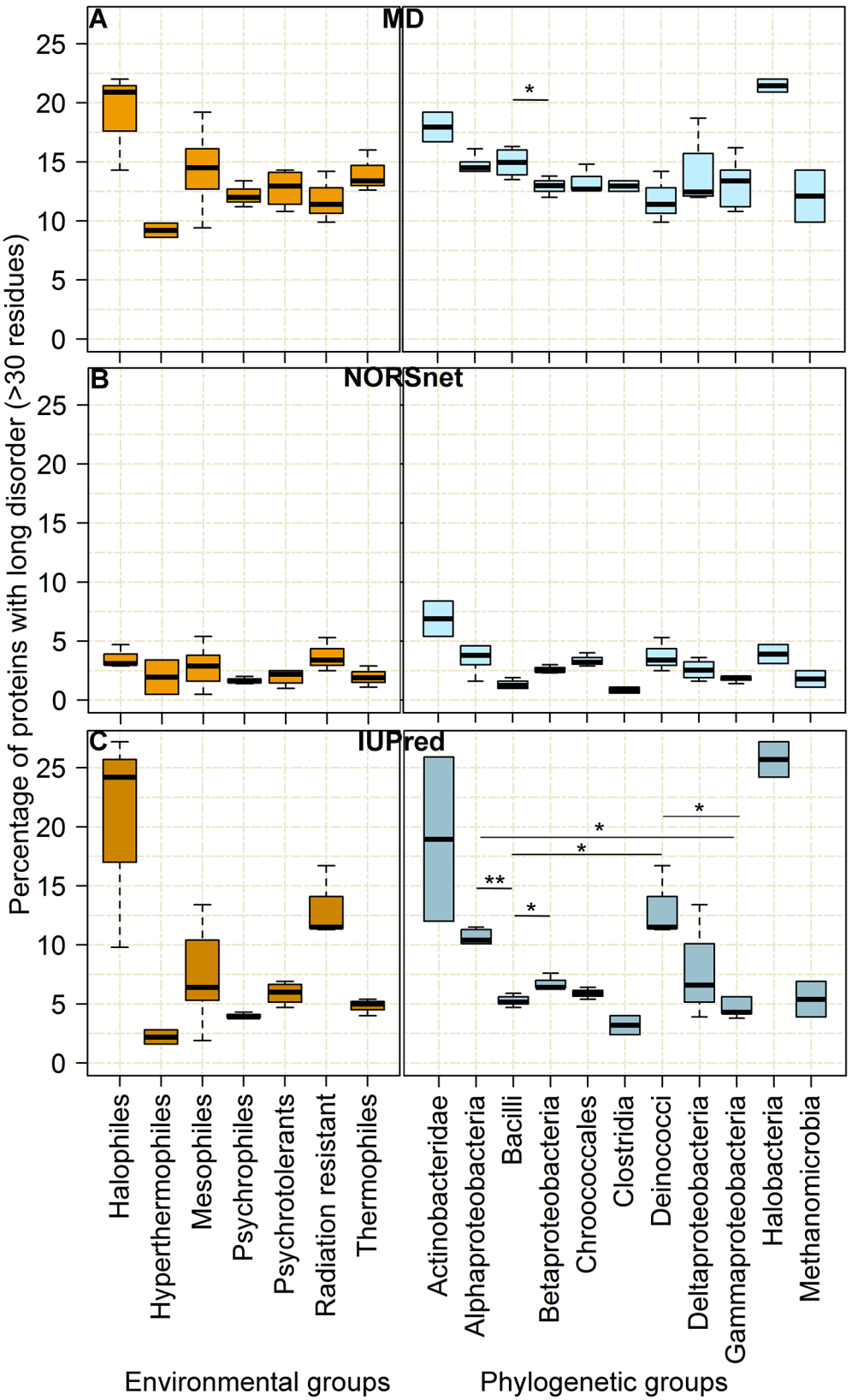

Fig 3. Protein disorder linked to habitat more than to phylogeny. The fractions of proteins with long disordered regions are predicted by two disorder predictor methods (MD in green bars and IUPred in red

bars). Eukaryotes are predicted with substantially more disorder than prokaryotes. Within the kingdoms predictions vary greatly: organisms in similar habitats tend to resemble each other in terms of disorder more than they resemble their closest phylogenetic relatives. (A) Hyperthermophilic archaea (dark red) are more ordered than their phylogenetic neighbors; halophilic archaea are more disordered (green). (B) Halophilic bacteria also appear more disordered than their relatives. (C) The bacterial thermophile (red) also has less disorder than its relatives. Other extreme organisms included: psychrophile (blue), psychrotolerant (light blue), radiation resistant (purple) and alkalophile (pink). We could also find organisms with relative high/low disorder content explainable separately.

doi:10.1371/journal.pone.0133990.g003

Conclusions

Extremophiles thrive in environments with extreme conditions such as high salt, exceptionally low or high temperatures and high radiation. We compared organisms through a quite simple criterion, namely the percentage of proteins for which at least one long region of disorder was predicted by 3x4 approaches to predict disorder (three methods, four thresholds). We analyzed protein disorder for several prokaryotic extremophiles and their closest phylogenetic relatives. We found protein disorder to be more reflective of habitat than of the evolutionary relation. This suggested that disordered regions might help crucially in adapting to challenging environments. For example, halophiles appeared to have significantly more protein disorder than their mesophilic relatives suggesting that protein disorder might compensate for the osmotic stress in extremely salty environments. Our data also indicated that the protein disorder differences between habitats depend less on the features of the corresponding taxonomic branch. For instance, both halophilic bacterial and halophilic archaeal proteomes were predicted with more disorder than their taxonomic neighbors. Correspondingly, hyperthermophiles appeared to have less disorder than their mesophilic taxonomic relatives. Finally, we investigated how disordered regions might contribute to environmental adaptation. Comparing the homologues between two extremophiles from cold and heat, we established that more often than expected by chance, disordered regions were found in the cold than in the heat. Largely, it appeared that the level of disorder was rather affected by many small than by few big changes. Overall, protein disorder appeared as a possible building block to bring about evolutionary changes such as the adaptation to different habitats.

Supporting Information

S1 Fig. Processing steps for "completely disordered" approach.
(PDF)

S2 Fig. Flowchart of statistical steps.
(PDF)

S3 Fig. Distribution of disorder content in different organisms for %long30.
(PDF)

S4 Fig. Distribution of disorder content in different organisms for %long50.
(PDF)

S5 Fig. Distribution of disorder content in different organisms for %long80.
(PDF)

S6 Fig. Graphical representation for amino acid abundance in different extreme organisms using Z-score.
(PDF)

S7 Fig. Protein disorder content by environment or phylogeny for %long50.
(PDF)

S8 Fig. Protein disorder content by environment or phylogeny for %long80.
(PDF)

S1 Table. List of organisms grouped after environmental conditions.
(PDF)

S2 Table. Z-score for protein disorder abundance for disorder regions > 30 residues.
(PDF)

S3 Table. Z-score for protein disorder abundance for disorder regions > 50 residues.
(PDF)

S4 Table. Z-score for protein disorder abundance for disorder regions > 80 residues.
(PDF)

S5 Table. Z-score for protein disorder abundance for "completely disordered" proteins.
(PDF)

S6 Table. Protein disorder abundance for disorder regions > 30 residues.
(PDF)

S7 Table. Protein disorder abundance for disorder regions > 50 residues.
(PDF)

S8 Table. Protein disorder abundance for disorder regions > 80 residues.
(PDF)

S9 Table. Protein disorder abundance for completely disordered proteins.
(PDF)

S10 Table. Overlap in protein disorder between a hyperthermophile and a mesophile.
(PDF)

S11 Table. Overlap in protein disorder between a psychrophile and a mesophile.
(PDF)

S12 Table. Relation protein disorder vs. ordered for homologue proteins in two extreme organisms.
(PDF)

S13 Table. Amino acid distribution on different groups of extreme organisms.
(PDF)

S14 Table. List of organisms grouped after taxonomical classification.
(PDF)

S15 Table. Test of equality of variances and medians of the groups for MD predictions (% long30/50/80).
(PDF)

S16 Table. Test of equality of variances and medians of the groups for NORSnet predictions (%long30/50/80).
(PDF)

S17 Table. Test of equality of variances and medians of the groups for IUPred predictions (%long30/50/80).
(PDF)

S18 Table. Test of equality of variances and medians of the groups for algorithm completely disordered.
(PDF)

S19 Table. Test of equality of variances and medians of the groups for Z-scores of MD predictions (%long30/50/80).
(PDF)

S20 Table. Test of equality of variances and medians of the groups for Z-scores of NORSnet predictions (%long30/50/80).
(PDF)

S21 Table. Test of equality of variances and medians of the groups for Z-scores of IUPred predictions (%long30/50/80).
(PDF)

S22 Table. Test of equality of variances and medians of the groups for Z-scores of algorithm completely disordered.
(PDF)

Acknowledgments

Thanks to Tim Karl and Laszlo Kajan (TUM) for invaluable help with hardware and software; to Marlena Drabik (TUM) for administrative support; to Christian Schaefer (Global Data Scientist at Allianz), Arthur Dong (TUM) and Edda Kloppmann (TUM) for helpful comments on the manuscript. Particular thanks to the editor and the three anonymous reviewers for an unprecedented amount of patience and help! More generally, thanks to all who deposit their experimental data in public databases, and to those who maintain these databases.

Author Contributions

Conceived and designed the experiments: AS EV BR. Performed the experiments: EV AS. Analyzed the data: EV BR AS. Contributed reagents/materials/analysis tools: AS BR EV. Wrote the paper: EV BR AS. Pictures: EV BR.

References

1. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder—a breakthrough invention of evolution? Curr Opin Struct Biol. 2011; 21(3):412–8. Epub 2011/04/26. doi: S0959-440X(11)00066-2 [pii] doi: 10.1016/j.sbi.2011.03.014 PMID: 21514145.
2. Wright PE, Dyson HJ. Linking folding and binding. Curr Opin Struct Biol. 2009; 19(1):31–8. Epub 2009/01/23. doi: S0959-440X(08)00179-6 [pii] doi: 10.1016/j.sbi.2008.12.003 PMID: 19157855; PubMed Central PMCID: PMC2675572.
3. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit. 2005; 18(5):343–84. Epub 2005/08/12. doi: 10.1002/jmr.747 PMID: 16094605.
4. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005; 272(20):5129–48. Epub 2005/10/13. doi: EJB4948 [pii] doi: 10.1111/j.1742-4658.2005.04948.x PMID: 16218947.
5. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 2005; 579(15):3346–54. Epub 2005/06/10. doi: S0014-5793(05)00424-2 [pii] doi: 10.1016/j.febslet.2005.03.072 PMID: 15943980.

6. Schlessinger A, Punta M, Rost B. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics. 2007; 23(18):2376–84. Epub 2007/08/22. doi: btm349 [pii] doi: 10.1093/bioinformatics/btm349 PMID: 17709338.
7. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008; 18(6):756–64. Epub 2008/10/28. doi: S0959-440X(08)00151-6 [pii] doi: 10.1016/j.sbi.2008.10.002 PMID: 18952168.
8. Singh GP, Dash D. Intrinsic disorder in yeast transcriptional regulatory network. Proteins. 2007; 68 (3):602–5. Epub 2007/05/19. doi: 10.1002/prot.21497 PMID: 17510967.
9. Fuxreiter M, Tompa P, Simon I, Uversky VN, Hansen JC, Asturias FJ. Malleable machines take shape in eukaryotic transcriptional regulation. Nat Chem Biol. 2008; 4(12):728–37. Epub 2008/11/15. doi: nchembio.127 [pii] doi: 10.1038/nchembio.127 PMID: 19008886; PubMed Central PMCID: PMC2921704.
10. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol. 2002; 322(1):53–64. Epub 2002/09/07. doi: S0022283602007362 [pii]. PMID: 12215414.
11. Devos D, Dokudovskaya S, Williams R, Alber F, Eswar N, Chait BT, et al. Simple fold composition and modular architecture of the nuclear pore complex. Proc Natl Acad Sci U S A. 2006; 103(7):2172–7. Epub 2006/02/08. doi: 0506345103 [pii] doi: 10.1073/pnas.0506345103 PMID: 16461911; PubMed Central PMCID: PMC1413685.
12. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000; 11:161–71. Epub 2001/11/09. PMID: 11700597.
13. Esnouf RM, Hamer R, Sussman JL, Silman I, Trudgian D, Yang ZR, et al. Honing the in silico toolkit for detecting protein disorder. Acta Crystallogr D Biol Crystallogr. 2006; 62(Pt 10):1260–6. Epub 2006/09/27. doi: S0907444906033580 [pii] doi: 10.1107/S0907444906033580 PMID: 17001103.
14. Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, et al. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 2011; 12(2):R14. doi: 10.1186/gb-2011-12-2-r14 PMID: 21324131; PubMed Central PMCID: PMC3188796.
15. Mohan A, Sullivan WJ Jr, Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol Biosyst. 2008; 4(4):328–40. Epub 2008/03/21. doi: 10.1039/b719168e PMID: 18354786.
16. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. Biochemistry and cell biology = Biochimie et biologie cellulaire. 2010; 88(2):167–74. doi: 10.1139/o09-163 PMID: 20453919.
17. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature. 2002; 420(6912):218–23. Epub 2002/11/15. doi: 10.1038/nature01256 nature01256 [pii]. PMID: 12432406.
18. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM. Protein interaction networks from yeast to human. Curr Opin Struct Biol. 2004; 14(3):292–9. Epub 2004/06/15. doi: 10.1016/j.sbi.2004.05.003 S0959440X04000776 [pii]. PMID: 15193308.
19. Montanari F, Shields DC, Khaldi N. Differences in the number of intrinsically disordered regions between yeast duplicated proteins, and their relationship with functional divergence. PLoS One. 2011; 6(9):e24989. doi: 10.1371/journal.pone.0024989 PMID: 21949823; PubMed Central PMCID: PMC3174238.
20. van der Lee R, Lang B, Kruse K, Gsponer J, Sanchez de Groot N, Huynen MA, et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell reports. 2014; 8(6):1832–44. doi: 10.1016/j.celrep.2014.07.055 PMID: 25220455.
21. Tompa P, Prilusky J, Silman I, Sussman JL. Structural disorder serves as a weak signal for intracellular protein degradation. Proteins. 2008; 71(2):903–9. doi: 10.1002/prot.21773 PMID: 18004785.
22. Gsponer J, Futschik ME, Teichmann SA, Babu MM. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science. 2008; 322(5906):1365–8. doi: 10.1126/science.1163581 PMID: 19039133; PubMed Central PMCID: PMC2803065.
23. Liu J, Rost B. Comparing function and structure between entire proteomes. Protein Sci. 2001; 10 (10):1970–9. Epub 2001/09/22. doi: 10.1110/ps.10101 PMID: 11567088; PubMed Central PMCID: PMC2374214.
24. Aravind L, Koonin EV. Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system. Genome Res. 2000; 10(8):1172–84. Epub 2000/08/25. PMID: 10958635; PubMed Central PMCID: PMC310937.
25. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helix prediction at 95% accuracy. Protein Science. 1995; 4:521–33. PMID: 7795533

26. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proceedings of the National Academy of Sciences. 1997; 94(22):11911–6.

27. Rost B. Did evolution leap to create the protein universe? Curr Opin Struct Biol. 2002; 12(3):409–16. Epub 2002/07/20. doi: S0959440X02003378 [pii]. PMID: 12127462.

28. Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, Sali A, et al. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. PLoS Biol. 2004; 2(12): e380. Epub 2004/11/04. doi: 10.1371/journal.pbio.0020380 PMID: 15523559; PubMed Central PMCID: PMC524472.

29. Petsko GA. Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat". Methods Enzymol. 2001; 334:469–78. Epub 2001/06/12. doi: S0076-6879(01) 34486-5 [pii]. PMID: 11398484.

30. Robinson-Rechavi M, Alibes A, Godzik A. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima. J Mol Biol. 2006; 356(2):547–57. Epub 2005/12/27. doi: S0022-2836(05)01479-8 [pii] doi: 10.1016/j.jmb.2005.11.065 PMID: 16375925.

31. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000; 13 (3):179–91. Epub 2000/04/25. PMID: 10775659.

32. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics. 2000; 1(1):76–88. Epub 2002/01/17. doi: 10.1007/s101420000003 PMID: 11793224.

33. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. Proteins. 2004; 54(1):20–40. Epub 2004/01/06. doi: 10.1002/prot.10559 PMID: 14705021.

34. D'Amico S, Collins T, Marx JC, Feller G, Gerday C. Psychrophilic microorganisms: challenges for life. EMBO Rep. 2006; 7(4):385–9. Epub 2006/04/06. doi: 7400662 [pii] doi: 10.1038/sj.embor.7400662 PMID: 16585939; PubMed Central PMCID: PMC1456908.

35. Paul S, Bag SK, Das S, Harvill ET, Dutta C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 2008; 9(4):R70. Epub 2008/04/10. doi: gb-2008-9-4-r70 [pii] doi: 10.1186/gb-2008-9-4-r70 PMID: 18397532; PubMed Central PMCID: PMC2643941.

36. Consortium TU. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2011; 40(Database issue):D71–5. Epub 2011/11/22. doi: gkr981 [pii] doi: 10.1093/nar/gkr981 PMID: 22102590; PubMed Central PMCID: PMC3245120.

37. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40(Database issue):D571–9. Epub 2011/12/03. doi: gkr1100 [pii] doi: 10.1093/nar/gkr1100 PMID: 22135293; PubMed Central PMCID: PMC3245063.

38. Mancinelli LJRRL. Life in extreme environments. Nature 2001; 409: 1092–101. doi: 10.1038/35059215 PMID: 11234023

39. Morita RY. Biological limits of temperature and pressure. Orig Life. 1980; 10(3):215–22. Epub 1980/09/01. PMID: 7413183.

40. Morita RY. Psychrophilic bacteria. Bacteriol Rev. 1975; 39(2):144–67. Epub 1975/06/01. PMID: 1095004; PubMed Central PMCID: PMC413900.

41. Schlessinger A, Liu J, Rost B. Natively unstructured loops differ from other loops. PLoS Comput Biol. 2007; 3(7):e140. Epub 2007/07/31. doi: 06-PLCB-RA-0416 [pii] doi: 10.1371/journal.pcbi.0030140 PMID: 17658943; PubMed Central PMCID: PMC1924875.

42. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. PLoS One. 2009; 4(2):e4433. Epub 2009/02/12. doi: 10.1371/journal.pone.0004433 PMID: 19209228; PubMed Central PMCID: PMC2635965.

43. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005; 21 (16):3433–4. Epub 2005/06/16. doi: bti541 [pii] doi: 10.1093/bioinformatics/bti541 PMID: 15955779.

44. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005; 347(4):827–39. Epub 2005/03/17. doi: S0022-2836(05)00129-4 [pii] doi: 10.1016/j.jmb.2005.01.071 PMID: 15769473.

45. Burra PV, Kalmar L, Tompa P. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PLoS One. 2010; 5(8):e12069. doi: 10.1371/journal.pone.0012069 PMID: 20711457; PubMed Central PMCID: PMC2920320.

46. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SC. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics. 2014. doi: 10.1093/bioinformatics/btu625 PMID: 25246432.

47. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007; 23(1):127–8. Epub 2006/10/20. doi: btl529 [pii] doi: 10.1093/bioinformatics/btl529 PMID: 17050570.

48. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 2011; 39(Web Server issue):W475–8. Epub 2011/04/08. doi: gkr201 [pii] doi: 10.1093/nar/gkr201 PMID: 21470960; PubMed Central PMCID: PMC3125724.

49. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2009; 37(Database issue):D26–31. doi: 10.1093/nar/gkn723 PMID: 18940867; PubMed Central PMCID: PMC2686462.

50. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2009; 37(Database issue):D5–15. doi: 10.1093/nar/gkn741 PMID: 18940862; PubMed Central PMCID: PMC2686545.

51. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012; 40(Database issue):D136–43. Epub 2011/12/06. doi: gkr1178 [pii] doi: 10.1093/nar/gkr1178 PMID: 22139910; PubMed Central PMCID: PMC3245000.

52. Usui K, Katayama S, Kanamori-Katayama M, Ogawa C, Kai C, Okada M, et al. Protein-protein interactions of the hyperthermophilic archaeon Pyrococcus horikoshii OT3. Genome Biol. 2005; 6(12):R98. Epub 2005/12/17. doi: gb-2005-6-12-r98 [pii] doi: 10.1186/gb-2005-6-12-r98 PMID: 16356270; PubMed Central PMCID: PMC1414084.

53. Methe BA, Nelson KE, Deming JW, Momen B, Melamud E, Zhang X, et al. The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. Proc Natl Acad Sci U S A. 2005; 102(31):10913–8. Epub 2005/07/27. doi: 0504766102 [pii] doi: 10.1073/pnas.0504766102 PMID: 16043709; PubMed Central PMCID: PMC1180510.

54. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25:3389–402. PMID: 9254694

55. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991; 9(1):56–68. Epub 1991/01/01. doi: 10.1002/prot.340090107 PMID: 2017436.

56. Abraham WR, Strompl C, Meyer H, Lindholst S, Moore ER, Christ R, et al. Phylogeny and polyphasic taxonomy of Caulobacter species. Proposal of Maricaulis gen. nov. with Maricaulis maris (Poindexter) comb. nov. as the type species, and emended description of the genera Brevundimonas and Caulobacter. Int J Syst Bacteriol. 1999; 49 Pt 3:1053–73. Epub 1999/07/30. PMID: 10425763.

57. Mika S, Rost B. UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res. 2003; 31(13):3789–91. Epub 2003/06/26. PMID: 12824419; PubMed Central PMCID: PMC169026.

58. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol. 2000; 18(6):609–13. Epub 2000/06/03. doi: 10.1038/76443 PMID: 10835597.

59. Natale DA, Galperin MY, Tatusov RL, Koonin EV. Using the COG database to improve gene recognition in complete genomes. Genetica. 2000; 108(1):9–17. Epub 2001/01/06. PMID: 11145426.

60. Chan Y, Walmsley RP. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. Phys Ther. 1997; 77 (12):1755–62. PMID: 9413454.

61. Fan C, Zhang D. A note on power and sample size calculations for the Kruskal-Wallis test for ordered categorical data. Journal of biopharmaceutical statistics. 2012; 22(6):1162–73. doi: 10.1080/10543406.2011.578313 PMID: 23075015.

62. Fitzgerald S, Dimitrov D, Rumrill P. The basics of nonparametric statistics. Work. 2001; 16(3):287–92. PMID: 12441458.

63. Shott S. Nonparametric statistics. Journal of the American Veterinary Medical Association. 1991; 198 (7):1126–8. PMID: 2045326.

64. Wolfe MHaDA. Nonparametric Statistical Methods. New York: John Wiley & Sons. 1973:27–33.

65. Wu P, Han Y, Chen T, Tu XM. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. Statistics in medicine. 2014; 33(8):1261–71. doi: 10.1002/sim.6026 PMID: 24132928.

66. Fox J,Weisberg H S. An R and S Plus Companion to Applied Regression. Second Edition S, editor2011.

67. Fox J. Applied Regression Analysis and Generalized Linear Models. Sage SE, editor 2008.

68. Team RC. R: A Language and Environment for Statistical Computing. In: Computing RFfS, editor. Vienna, Austria2013.

69. Goo YA, Yi EC, Baliga NS, Tao WA, Pan M, Aebersold R, et al. Proteomic analysis of an extreme halophilic archaeon, Halobacterium sp. NRC-1. Mol Cell Proteomics. 2003; 2(8):506–24. Epub 2003/07/23. doi: 10.1074/mcp.M300044-MCP200 M300044-MCP200 [pii]. PMID: 12872007.

70. Oren A, Ginzburg M, Ginzburg BZ, Hochstein LI, Volcani BE. Haloarcula marismortui (Volcani) sp. nov., nom. rev., an extremely halophilic bacterium from the Dead Sea. Int J Syst Bacteriol. 1990; 40 (2):209–10. Epub 1990/04/01. PMID: 11536469.

71. Xia Q, Hendrickson EL, Zhang Y, Wang T, Taub F, Moore BC, et al. Quantitative proteomics of the archaeon Methanococcus maripaludis validated by microarray analysis and real time PCR. Mol Cell Proteomics. 2006; 5(5):868–81. Epub 2006/02/21. doi: M500369-MCP200 [pii] doi: 10.1074/mcp.M500369-MCP200 PMID: 16489187; PubMed Central PMCID: PMC2655211.

72. Ohta Y, Hatada Y, Nogi Y, Li Z, Ito S, Horikoshi K. Cloning, expression, and characterization of a glycoside hydrolase family 86 beta-agarase from a deep-sea Microbulbifer-like isolate. Appl Microbiol Biotechnol. 2004; 66(3):266–75. Epub 2004/10/19. doi: 10.1007/s00253-004-1757-5 PMID: 15490156.

73. Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. 2013; 22(6):693–724. doi: 10.1002/pro.2261 PMID: 23553817; PubMed Central PMCID: PMC3690711.

74. Gunbin KV, Afonnikov DA, Kolchanov NA. Molecular evolution of the hyperthermophilic archaea of the Pyrococcus genus: analysis of adaptation to different environmental conditions. BMC Genomics. 2009; 10:639. Epub 2010/01/01. doi: 1471-2164-10-639 [pii] doi: 10.1186/1471-2164-10-639 PMID: 20042074; PubMed Central PMCID: PMC2816203.

75. Fukuhara H, Kifusa M, Watanabe M, Terada A, Honda T, Numata T, et al. A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3. Biochem Biophys Res Commun. 2006; 343(3):956–64. Epub 2006/04/01. doi: S0006-291X(06)00489-X [pii] doi: 10.1016/j.bbrc.2006.02.192 PMID: 16574071.

76. Sako Y, Nomura N, Uchida A, Ishida Y, Morii H, Koga Y, et al. Aeropyrum pernix gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100 degrees C. Int J Syst Bacteriol. 1996; 46(4):1070–7. Epub 1996/10/01. PMID: 8863437.

77. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, et al. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. DNA Res. 1999; 6 (2):83–101, 45–52. Epub 1999/06/26. PMID: 10382966.

78. Milek I, Cigic B, Skrt M, Kaletunc G, Ulrih NP. Optimization of growth for the hyperthermophilic archaeon Aeropyrum pernix on a small-batch scale. Can J Microbiol. 2005; 51(9):805–9. Epub 2006/01/05. doi: w05-060 [pii] doi: 10.1139/w05-060 PMID: 16391661.

79. Otgonbayar GE, Eom HJ, Kim BS, Ko JH, Han NS. Mannitol production by Leuconostoc citreum KACC 91348P isolated from Kimchi. J Microbiol Biotechnol. 21(9):968–71. Epub 2011/09/29. doi: JMB021-09-11 [pii]. PMID: 21952374.

80. Tantos A, Friedrich P, Tompa P. Cold stability of intrinsically disordered proteins. FEBS Lett. 2009; 583(2):465–9. Epub 2009/01/06. doi: S0014-5793(08)01040-5 [pii] doi: 10.1016/j.febslet.2008.12.054 PMID: 19121309.

81. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, et al. Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev. 2001; 65(1):44–79. Epub 2001/03/10. doi: 10.1128/MMBR.65.1.44–79.2001 PMID: 11238985; PubMed Central PMCID: PMC99018.

82. Cox MM, Battista JR. Deinococcus radiodurans—the consummate survivor. Nat Rev Microbiol. 2005; 3(11):882–92. Epub 2005/11/02. doi: nrmicro1264 [pii] doi: 10.1038/nrmicro1264 PMID: 16261171.

83. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, et al. Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res. 2000; 28(21):4317–31. Epub 2000/11/01. PMID: 11058132; PubMed Central PMCID: PMC113120.

84. Heidelberg JF, Paulsen IT, Nelson KE, Gaidos EJ, Nelson WC, Read TD, et al. Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. Nat Biotechnol. 2002; 20 (11):1118–23. Epub 2002/10/09. doi: 10.1038/nbt749 nbt749 [pii]. PMID: 12368813.

85. Tsang PH, Li G, Brun YV, Freund LB, Tang JX. Adhesion of single bacterial cells in the micronewton range. Proc Natl Acad Sci U S A. 2006; 103(15):5764–8. Epub 2006/04/06. doi: 10.1073/pnas.0601705103 PMID: 16585522; PubMed Central PMCID: PMC1458647.

86. Hopkin M. Bacterium makes nature's strongest glue. Nature. 2006.

87. Christie-Oleza JA, Miotello G, Armengaud J. High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics. 13:73. Epub 2012/02/18. doi: 1471-2164-13-73 [pii] doi: 10.1186/1471-2164-13-73 PMID: 22336032; PubMed Central PMCID: PMC3305630.

88. Yi H, Lim YW, Chun J. Taxonomic evaluation of the genera Ruegeria and Silicibacter: a proposal to transfer the genus Silicibacter Petursdottir and Kristjansson 1999 to the genus Ruegeria Uchino et al. 1999. IJSEM. 2007; 57(4):815–9. doi: 10.1099/ijs.0.64568–0

89. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12(2):85–94. Epub 1999/04/09. PMID: 10195279.

90. Lee DY, Kim KA, Yu YG, Kim KS. Substitution of aspartic acid with glutamic acid increases the unfolding transition temperature of a protein. Biochem Biophys Res Commun. 2004; 320(3):900–6. Epub 2004/07/09. doi: 10.1016/j.bbrc.2004.06.031 S0006291X0401294X [pii]. PMID: 15240133.

91. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A. 1984; 81(1):140–4. Epub 1984/01/01. PMID: 6582470; PubMed Central PMCID: PMC344626.

92. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157(1):105–32. Epub 1982/05/05. doi: 0022-2836(82)90515-0 [pii]. PMID: 7108955.

93. Janin J. Surface and inside volumes in globular proteins. Nature. 1979; 277(5696):491–2. Epub 1979/02/08. PMID: 763335.

94. Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Molecular biology and evolution. 2013; 30(12):2645–53. doi: 10.1093/molbev/mst157 PMID: 24037790.

95. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteome Res. 2006; 5(4):879–87. doi: 10.1021/pr060048x PMID: 16602695; PubMed Central PMCID: PMC2543136.

96. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. J Proteome Res. 2006; 5(4):888–98. doi: 10.1021/pr060049p PMID: 16602696; PubMed Central PMCID: PMC2533134.

97. Denning DP, Rexach MF. Rapid evolution exposes the boundaries of domain structure and function in natively unfolded FG nucleoporins. Mol Cell Proteomics. 2007; 6(2):272–82. doi: 10.1074/mcp.M600309-MCP200 PMID: 17079785.

98. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. Journal of molecular evolution. 2002; 55(1):104–10. doi: 10.1007/s00239-001-2309-6 PMID: 12165847.

99. Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of molecular evolution. 2002; 55(3):260–4. doi: 10.1007/s00239-002-2323-3 PMID: 12187379.

100. Pavlovic-Lazetic GM, Mitic NS, Kovacevic JJ, Obradovic Z, Malkov SN, Beljanski MV. Bioinformatics analysis of disordered proteins in prokaryotes. BMC Bioinformatics. 2011; 12:66. doi: 10.1186/1471-2105-12-66 PMID: 21366926; PubMed Central PMCID: PMC3062596.

101. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. Journal of molecular graphics & modelling. 2001; 19(1):26–59. PMID: 11381529.

102. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993; 232(2):584–99. doi: 10.1006/jmbi.1993.1413 PMID: 8345525.

RESEARCH ARTICLE

Protein disorder reduced in *Saccharomyces cerevisiae* to survive heat shock [version 1; referees: 2 approved, 1 approved with reservations]

Esmeralda Vicedo[1,3], Zofia Gasik[1,2], Yu-An Dong[1,4], Tatyana Goldberg[1], Burkhard Rost[1,5,6]

[1]Department of Informatics, Bioinformatics & Computational Biology, TUM, Munich, Germany
[2]Graduate School of Information Science in Health, TUM, Munich, Germany
[3]Institute of Experimental Physics, Division of Biophysics, University of Warsaw, Warsaw, Poland
[4]Institute of Systems Biology, Shanghai University, Shanghai, China
[5]Institute of Advanced Study, TUM, Munich, Germany
[6]Institute for Food and Plant Sciences WZW, TUM, Freising, Germany

v1 **First published:** 06 Nov 2015, **4**:1222 (doi: 10.12688/f1000research.7178.1)
Latest published: 06 Nov 2015, **4**:1222 (doi: 10.12688/f1000research.7178.1)

Abstract

Recent experiments established that a culture of *Saccharomyces cerevisiae* (baker's yeast) survives sudden high temperatures by specifically duplicating the entire chromosome III and two chromosomal fragments (from IV and XII). Heat shock proteins (HSPs) are not significantly over-abundant in the duplication. In contrast, we suggest a simple algorithm to "*postdict*" the experimental results: Find a small enough chromosome with minimal protein disorder and duplicate this region. This algorithm largely explains all observed duplications. In particular, all regions duplicated in the experiment reduced the overall content of protein disorder. The differential analysis of the functional makeup of the duplication remained inconclusive. Gene Ontology (GO) enrichment suggested over-representation in processes related to reproduction and nutrient uptake. Analyzing the protein-protein interaction network (PPI) revealed that few network-central proteins were duplicated. The predictive hypothesis hinges upon the concept of reducing proteins with long regions of disorder in order to become less sensitive to heat shock attack.

Open Peer Review

Referee Status: ✓ ? ✓

	Invited Referees 1	2	3
version 1 published 06 Nov 2015	✓ report	? report	✓ report

1 **Melchor Sanchez-Martinez**, Mind the Byte Spain

2 **Paul Pavlidis**, University of British Columbia Canada

3 **Anuj Kumar**, University of Michigan USA

Discuss this article

Comments (0)

Corresponding author: Esmeralda Vicedo (vicedo@rostlab.org)

How to cite this article: Vicedo E, Gasik Z, Dong YA *et al.* **Protein disorder reduced in *Saccharomyces cerevisiae* to survive heat shock [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2015, **4**:1222 (doi: 10.12688/f1000research.7178.1)

Grant information: This work, and all authors were supported by the German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 06 Nov 2015, **4**:1222 (doi: 10.12688/f1000research.7178.1)

Introduction

Saccharomyces cerevisiae (baker's yeast; for simplicity we mostly use yeast) was the first completely sequenced eukaryote[1]. Being simple to handle and manipulate has rendered yeast a preferred model organism for genetics, biochemistry and systems biology[2–4]. It grows optimally within a narrow temperature range but tolerates moderate deviations, some of which impinge upon cell structure and function, often through rapid physiological adaptations. One such adaptation mechanism is the duplication of the whole genome or particular chromosomes (aneuploidy)[5–7] that contain the genes necessary to rapidly cope with specific adverse conditions over the course of several generations of evolving yeast[8–14]. Such evolutionary adaptations imbalance the genome[15], destabilize reactions or pathways[16,17], and cost energy[18,19]. Aneuploidy, therefore, is a transient solution. Over many generations exposed to the same adverse conditions refined specific and less expensive solutions replace aneuploidy[20]. Yeast cells can adapt to high-temperature stress by repeatedly duplicating chromosome III along with two other fragments (from chrIV and chrXII)[20]. Why specifically copy these regions? Can particular biophysical features and/or functions of the proteins encoded in these regions explain the choice?

One simple biophysical feature is protein disorder: most proteins adopt well-defined three-dimensional (3D) structures[21–24], *i.e.* will largely remain identical at different times. In contrast, *disordered regions* do not adopt well-defined 3D structures in isolation[25], *i.e.* without binding substrates they will look very different at different time points. Proteins with long disordered regions encompass some unique biophysical characteristics[26–34]. Such regions are so difficult to characterize experimentally that there is no good experimental data set proxy for "all proteins with long regions of disorder in yeast". In contrast, acceptably accurate computational predictions are available for entire proteomes[30,35,36]. Protein disorder seems one means for prokaryotes to adopt to extreme environments, e.g. halophiles have more proteins with long disorder than their closest phylogenetic relatives, while thermophiles tend to have fewer[37]. Here, we hypothesized a similar effect to govern the response to high temperature-related duplication in yeast, namely that chromosomal regions duplicated under high temperature are depleted of proteins with long disorder.

Methods

Data

We downloaded the yeast (*S. cerevisiae*) proteome from UniProt (proteome ID: UP000002311)[38] as fasta files including only the reviewed proteins (UniProtKB/Swiss-Prot). Removal of duplicates applying the method Uniqueprot2[39] (with 100% pairwise sequence identity, keeping the longer sequence) left 5667 proteins (Table S1A). We considered the 16 nuclear chromosomes (matched through http://www.yeastgenome.org[40], the numbers of proteins per chromosome are given in Figure S1B). The yeastgenome.org resource also provided the annotations of heat-shock response proteins (HSR). Proteins known to interact with HSR proteins augmented this set of HSRs in the following way.

BioGRID (version 3.1.86) provided the data for experimental protein-protein interactions (PPIs) in yeast. After filtering out redundancy (a-b and b-a counted only once) and excluding self-interactions (a-a), we based all subsequent analysis on the single largest connected component of the network. We focused on the most basic network features that allow the comparison and characterization of complex networks. The most elementary characteristic of a node is its degree or connectivity, defined as the number of interactions for a node (here protein), i.e. the number of interactions one protein has with all others. Another important parameter is the betweenness, *i.e.* the fraction of shortest paths between all other nodes that has to go through a given node. Additionally, we monitored the average degree of neighbors, which depends on the number of nodes and links in the network. These three parameters measured the importance of each node within the network.

Disorder predictions

We applied methods capturing different "flavors" of protein disorder[29,41,42]. IUPred (version 1.0) is based on statistical contact potentials and exclusively uses single sequences[28,43]. MD (MetaDisorder)[42] combines different original prediction methods through machine learning (neural network) with evolutionary profiles and predictions of solvent accessibility and protein flexibility. To some extent disorder is a gradual phenomenon, i.e. proteins may have more or less disorder[44]. On the other hand, prediction methods distinguish between a 30-residue loop resembling "protein disorder" and another resembling a region with "regular structure"[29]. Thus, protein disorder seems more a binary feature (it is there or not, or present/absent) than a gradual one[25]. Unfortunately, no argument or data determines one single correct threshold for what constitutes present/absent for protein disorder. Typically, experts use a length threshold of the type: protein disorder is present when at least T consecutive residues in a protein are predicted to be disordered. If so, this protein is considered to contain a long region of disorder. More disorder in this model could imply, e.g. more than one region, or the entire protein. We analyzed many alternatives to choose the threshold for long disorder, and found most to be redundant. We included different views only if they provided relevant information. In particular, we largely focused on one threshold to define "long disorder": **%long30**, is the percentage of proteins with at least one region of ≥30 consecutive residues predicted as disordered (alternatives were: %long50 and %long80, i.e. with length thresholds at ≥50 and ≥80, and completely disordered implying no region of 30 consecutive residues without any disorder).

GO term enrichment

We applied BINGO (Biological Networks Gene Ontology[44], version 2.44) to identify the enrichment of GO (Gene Ontology)[45] terms in subsets of experimentally annotated proteins. We focused on "biological process" and "molecular function". For two sets of proteins with annotated biological functions (more precisely: GO numbers) BINGO estimates to which extent their annotations differ in a statistically significant way. We visualized BINGO results with Cytoscape[46] platform (version 2.8). Our analysis focused on the hypergeometric test in BINGO, which accurately estimates p-values as it tests without replacement. Following the common, procedure for BINGO, we considered p-values >0.05 as significant[46]. Testing multiple hypotheses may give many false positives (Type I error: incorrect rejection of true null hypothesis[47,48]). Using BINGO, we corrected for these through the Benjamini and Hochberg correction which provides strong control over the False Discovery Rate

(FDR, expected proportion of erroneous null hypothesis rejections among all rejections[48]).

Results and discussion

Duplications in response to high temperature reduce protein disorder

In response to high temperature yeast (*S. cerevisiae*) duplicates the entire chromosome III (for brevity we use *chrN* to denote 'yeast chromosome N' with N as Roman numerals following convention) and fragments from chrIV and chrXII[20]. The size of the 16 yeast chromosomes varies over six-fold (Table S1). The average protein length is similar between the 16 chromosomes (Figure S1, Table S1). The duplicated chrIII is the 3rd smallest with 183 genes, of which 153 are mapped and 132 constitute "verified ORFs". Fewer genes are encoded only by chrI with 90, and chrVI with 125 proteins (Table S1). The relatively small number of genes on chrIII was one reason for choosing it as the first fully synthesized functional yeast chromosome[49]. In contrast to protein length, the percentage of proteins with predicted long regions of disorder differed significantly between the 16 yeast chromosomes (Figure 1).

The least protein disorder was predicted for chrIII and chrX (Figure 1, Table S2). That means heat response duplicates one of the two chromosomes with the least disorder. In addition, the fragments of chrIV and chrXII that are duplicated along with the entire chrIII also clearly have less disorder than the chromosomes from which they were taken (Figure 1). This enhances the effect of protein disorder reduction in response to high temperatures.

The other low-disorder option is chrX: Why not duplicate chrX in response to high temperatures? ChrX is more than twice as large as chrIII (Figure S1). Thus duplicating chrX would "cost" twice as much. This might be prohibitive. ChrX might also not contain the cell activities important for coping with high temperature. Furthermore, as chrX and chrIII are similar in disorder content while chrX has twice the proteins of chrIII, the duplication of chrX would increase the overall level of proteins with disorder that might become unfolded and thereby "jam" cellular activity more than the duplication of chrIII.

Assume a certain amount of tolerable duplication were tolerable and that number were about 153 proteins (as for chrIII): where in the genome do we find a continuous stretch (within a chromosome) that has 153 proteins with the least disorder? Our results underscored the special role of chrIII (Figure S2): only 3% of all continuous genome fragments with 153 proteins have as little disorder as chrIII (corresponding numbers for chrX: 5%; 29-protein fragment from chrIV: 52%; 64-protein fragment from chrXII: 10%). These figures demonstrate that the duplication of chrIII might be THE optimal choice for a simple way to duplicate 153 proteins with as little disorder as possible.

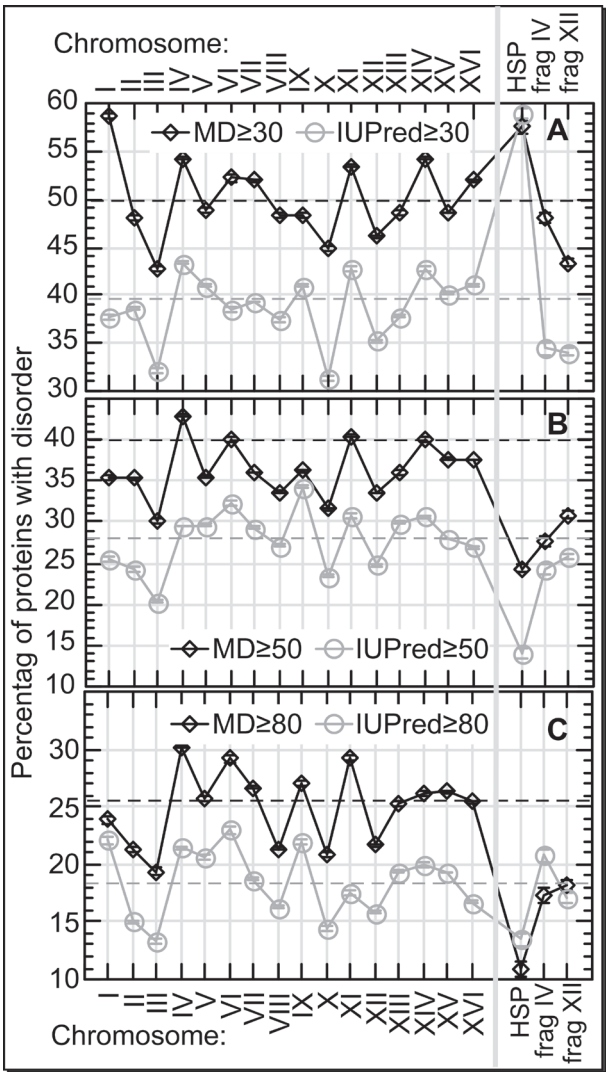

Figure 1. Protein disorder differs between yeast chromosomes. The composition of proteins with long regions of disorder (y-axes) differed significantly between the 16 chromosomes of *S. cerevisiae* (x-axes) and also for the set of HSPs. The three rightmost marks on the x-axes describe: HSPs and the disorder predictions for the HSR-related duplicated fragments on chromosome IV and chromosome XII (frag IV and frag XII). The differences were similar for two different prediction methods (MD in black, IUPred in light gray), and for different thresholds with respect to the minimal length of a disordered region (**A**: ≥30 consecutive residues predicted in disorder, **B**: ≥50, **C**: ≥80). Dashed horizontal lines mark the averages over all chromosomes. Error bars are too small to become visible on the scale chosen. The least disorder content was predicted for chromosome III and chromosome X. Overall, all duplications in response to heat shock treatment reduced the level of protein disorder in the offspring.

Heat-shock proteins do not explain the temperature-related duplication

Our results explained why duplicating 150–200 proteins from another chromosome might have been potentially more damaging than the duplication of chrIII in response to high temperatures. In other words, our model might suggest why the duplication of this particular region is better than other duplications. However, what is the selective benefit from the proteins on chrIII? We expected to find the answer to this question in proteins that actively help with coping with heat stress. The immediate suspects are heat-shock proteins (HSP) and the proteins known to interact with these HSPs (HSP-binders). The known HSPs and HSP-binders scatter over all 16 yeast chromosomes (Figure S3). All regions duplicated in response to heat shock contain only one known gene coding for HSPs (*HSP30*) and one known HSP-binder (*TAH1*). This implied that 1.3% of all known HSPs and HSP-binders were duplicated in an event that duplicated 0.5% of all genes, i.e. a 2.6-fold over-representation. This statistically insignificant the finding that fewer than 1 in 50 of all HSPs and HSP-binders are duplicated might still be scientifically significant if *HSP30* and *TAH1/HSP90* were outstandingly important proteins for the given conditions. However, this is not the case, at least not given what is currently known about *HSP30*. Furthermore, introducing an extra copy of *HSP30* into wild-type cells did not increase the ability of the cells to cope with high temperature (Dahan & Pilpel, personal communication).

The set of known HSPs (Figure S3) slightly changed expression levels in response to heat stress during the fixation of the trisomy[20] only slightly but almost all HSPs were significantly up-regulated (arrows in Figure S3) when the "refined descendants" replaced the trisomy[20]. This could imply that the duplicated genes are essential for survival under heat stress. Nevertheless, quite contrary to the naïve expectation, the HSPs and the HSP-binders by no means explained the heat-stress-specific duplications observed experimentally.

Incidentally, HSPs appeared particularly abundant in disorder regions of 30–50 consecutive residues (Figure 1A, in particular for IUPred). It has previously been argued that such disorder is required for proper function of HSPs[50]. In contrast, HSPs are depleted of longer disorder (>50; Figure 1B,C).

Overall, we argue that HSPs could have explained the duplication of many other chromosomes, possibly even better than that of chrIII. Therefore, this explanation is not specific. Thus, we conclude that the duplication of known HSPs and HSP-binding proteins did not explain why chrIII was specifically duplicated. Many HSPs and HSP-binders might remain unknown. However, we have no scientific ground to suspect that the fraction of the unknown HSPs differs between the chromosomes, i.e. that there are particular HSPs on chrIII that remain undiscovered.

GO terms enriched for growth and reproduction in heat stress-duplication

Are any other proteins on chrIII important for growth under high temperature? Simply scanning GO[45] annotations is insufficient: the question is not whether proteins on chrIII have certain functions, but whether these are overrepresented enough to explain why chrIII and not the other two small chromosomes (chrVI or chrI) are duplicated in response to high temperatures. In order to address this question, we need a GO term enrichment analysis of the duplicated regions[51].

Growth and reproduction might be considered as the most important cell activities in the sense that the organism must grow and proliferate (cells that fail are not observed) even under stress. The GO enrichment analysis seemed to confirm this expectation (Figure 2): the two most abundant GO terms in the heat stress-duplicated regions were those related to (i) sexual reproduction (Figure 2 and Figure S2; "conjugation with cellular fusion", "reproductive cellular process" and "response to pheromone") and to (ii) sugar transport (hexose transport process as well as mannose, fructose and glucose transmembrane transporter activity; Figure S4).

The major energy source of yeast is sugar, in particular hexose monosaccharides (C6H12O6; e.g. glucose, fructose, mannose). These nutrient sugars are imported into the cell through hexose transporters, which are encoded by *HXT* genes[52,53]. The *HXT* yeast genes on the duplicated fragment of chromosome IV (*HXT3*, *HXT6 and HXT7*) are almost five-fold over-represented with respect to random (yeast has 5667 N_{genY} genes, 243 N_{genD} are duplicated, 15 N_{genHXT} genes are in yeast; in a region with 243 N_{genD} genes we would find by chance 0.64 HXT genes in the duplicated regions $p_{chance}=N_{genHXT}*[N_{genD}/N_{genY}]$). Two HXT genes on the duplicated chrIV fragment (*HXT6* and *HXT7*) appear to encode high-affinity transporters required for growth at very low glucose concentrations (~0.1%[54]), *i.e*. these two would become particularly important when yeast is cultured under glucose limitation[54]. Interestingly, several works have detected duplication of these two genes (*HXT6* and *HXT7*) in yeast populations evolving under low nutrient availability[8,55]. These numbers suggest that heat stress also puts strain upon obtaining the energy needed for growth and reproduction.

Sexual reproduction also appeared crucial for the survival of yeast cultured under heat stress[56,57]. Seven of the ten molecular functions to be significantly overrepresented in the heat stress-duplicated chrIII (Table S3) by a standard GO-term enrichment analysis[51] are involved in reproduction. Three of these seven molecular functions are related specifically to sexual reproduction; the others pertain to general reproductive processes (Figure 2). In particular, the reproduction-related processes involve cell fusion (*FUS1* and *FIG2*[58–60]), pheromone response (*STE50* which is also required for optimal invasive growth and hyperosmotic stress signaling[61,62] and *NOT1* that is also involved in several RNA regulation levels[63]), nuclear fusion, chromosome disjunction, nuclear segregation after mating (*BIK1* which is involved in microtubule function during mitosis[64,65]), fusion of haploid nuclei during mating; *KAR4* or KARyogamy plays a critical role in the choreography of the mating response[66]), cytokenesis (division of cytoplasma and plasma membrane of a cell and its separation into two daughter cells which is also relevant for asexual mitotic growth: *CDC10*[67]), specification of the site where the daughter cell will form (relevant for budding and asexual growth, also referred to as axial bud selection) and in the developmental process in which the size of a cell is generated and organized (also referred to as morphogenesis: *CDC10*[67–69]). All these genes are also required for the correct

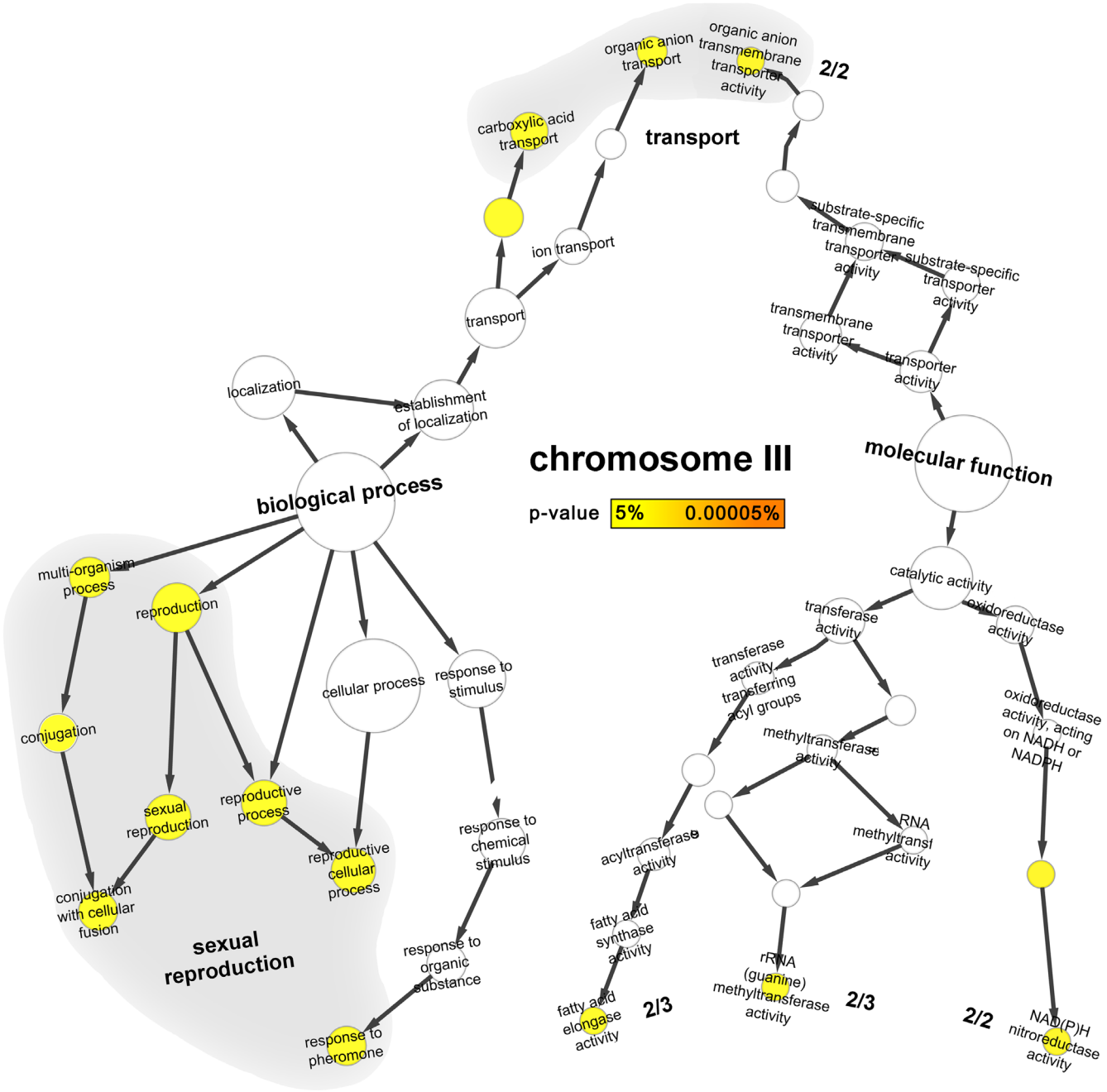

Figure 2. GO enrichment of sexual reproduction and nutrient uptake. The tree gives the complete set of all experimentally annotated GO-terms (Gene Ontology[45]) for any of the proteins on chromosome III that describe *biological process* (left branch) and *molecular function* (right branch). The enrichment analysis[51] describes how much chrIII GO-terms are enriched with respect to all other GO-terms from yeast: all terms marked by yellow circles are significantly enriched. Sexual reproduction (7 GO-terms on chrIII) and transport (carboxylic acid and organic anion 4 GO-terms on chrIII) mapped to most overrepresented GO terms on this chromosome.

localization of other proteins involved in cytokinesis and bud site selection[67,70–73]. Other important processes and activities overrepresented on chrIII are related to the avoidance of oxidative stress (e.g. carboxylic acid transport – Figure 2 - which may be important for the survival since during the vegetative asexual reproduction cells were exposed to oxidative stress) and NAD(P)H nitro-reductase activity (Figure 2). The only nitroreductase-related proteins in yeast – HBN1 and FRM2[74] - are only on chrIII (Table S4). The proteins involved in these two activities (carboxylic acid transport and NAD(P)H nitro-reductase activities) are also implicated in cellular detoxification[75], which is another task relevant for survival under stress.

All these data supported the view that chrIII is important for sexual reproduction. A seemingly convincing story, until we learned that the laboratory strains of yeast survived through asexual reproduction[20], *i.e.* apparently did not need what is so uniquely enriched in the heat stress duplication. The set of proteins known to be involved in reproduction on chrIII (Table S3) had more disorder than the average for chrIII (Figure S5). Some of these proteins with long disordered regions might not work correctly in heat.

Why duplicate proteins that fail? Not having found a convincing answer, we propose two conjectures: first, sexual reproduction might "frame" another cellular activity of the same protein that is more relevant to the growth conditions applied during the evolution in the laboratory experiment. For instance, *CDC10* is also required to maintain cell polarity (GO: 0030011), *BUD3* and *BUD5* are involved in axial cellular bud-site selection (GO: 0007120), *KCC4* a bud neck kinase involved in budding and cell bud growth (GO: 0007117) and *BIK1*, which is involved in microtubule function during mitosis. All of these activities are related to asexual reproduction. Our second proposition seems more far-fetched, namely that the set of proteins with the strongest GO-enrichment might have been duplicated coincidentally, *i.e.* the disorder-rich proteins related to sexual reproduction might have been duplicated because they happened to be on chrIII but not due their relevance for the survival in heat. If so, there must be something else we have not found yet on chrIII.

Several other processes were slightly enriched in the duplicated fragments with some relevance for yeast survival in heat but none of those gave a clear explanation (Figure 2): (i) fatty acid elongase, (ii) rRNA (guanine) methyltransferase, and (iii) the importin-alpha export receptor activities. We analyzed these in detail. (i) Fatty acid elongase: currently, only three proteins are known to be involved in lengthening fatty acids; two of those (ELO2 and APA1; Table S3) are on chrIII. Fatty acid elongases are involved in sphingolipid biosynthesis. The sphingolipids are components of the cellular membrane and bioactive signaling molecules that contribute to heat tolerance as they are directly involved in organizational cellular structures (e.g. cell membrane)[76]. (ii) rRNA methyltransferases: three yeast proteins are known to be involved in rRNA (guanine) methyltransferase activity; two of those (*BUD23* and *SPB1*) are on chrIII (Table S4). It is believed that the modification of ribonucleotides optimizes the rRNA structure and represents a way to expand the topological potentials of RNA molecules. It is possible that the loss of modification affects fine-tuning of ribosome function that could give rise to the pronounced cold-sensitivity[77]. (iii) Importin-alpha nuclear export: two yeast proteins contribute toward the importin-alpha export receptor activity; one of those (MSN5) is in the duplicated fragment of chromosome IV. *MSN5* knockout mutants show a variety of phenotypes, including carbon-source utilization, defects and sensitivity to high concentrations of ions, severe heat shock, and high pH[78]. Moreover, these mutants are partially sterile[78]. Therefore, this protein appears necessary for cell survival, especially under extreme conditions.

Only one cellular activity related to tRNA synthase appeared over-represented on the duplicated fragment of chrXII (DUS3 and DUS4 proteins; Table S7). In particular, to the tRNA dihydrouridine synthases, which are responsible for the reduction of the 5,6-double bond of a uridine residue on tRNA (one of the numerous modifications observed on tRNA cytoplasmatic[79]). However, this particular finding appeared less relevant since the corresponding fragment was only duplicated in one of four growth experiments in response to high temperatures[20].

One crucial limitation for any functional enrichment study remains the incomplete experimental annotation even for an organism as intensively studied as yeast. It may be that all our speculation above missed the real causation because the functions of the proteins that are really relevant remain uncharacterized. Therefore, we complemented our analysis with one aspect of function for which we have a complete prediction, namely the prediction of sub-cellular localization of all yeast proteins. The experimental localization annotations for yeast are still cover at most 70% of all proteins[80]. However, today's top prediction methods, such as LocTree3, are very reliable[80] and can make crucial differences for comparing 'complete' data sets[81]. We found nuclear proteins to be clearly depleted on chrIII (-4.6 percentage points with respect to the entire proteome; Figure S7A). Other abundant proteins found on chrIII were secreted (extra-cellular) or annotated as endoplasmic reticulum (ER) membrane proteins (each 3.2 percentage points higher than in the full yeast proteome). We also observed significantly more disorder in nuclear proteins (nuclear 77% vs. <40% for non-nuclear; Figure S8). This might explain the depletion of nuclear proteins on chrIII. While these findings were clear, they did not suggest a simple interpretation. The abundance of secreted proteins on chrIII (about 3.2 percentage points more on chrIII than in entire yeast; Figure S7A) implies that in the response to heat shock, more proteins are secreted into the 'hot' environment. Given the correlation between habitat and disorder[37], we expect that proteins are more likely to sustain high temperatures with less disorder. Unfortunately, a GO enrichment study of the secreted proteins also did not provide the answer we had been hoping for. However, the "secretome" alone could not explain the lower content of disordered proteins on chromosome III (disorder entire yeast-chrIII=50%-43%=7%>3% for secretome; Table 1 and Figure S7A).

Proteins from chrIII less implied in overall PPI network

As proteins cannot be understood without also considering their networks of interaction, we compared the network of experimentally characterized PPIs between the entire yeast and those fragments that are duplicated in heat evolving populations. As for the differential analysis of any experimental annotation, the limitation of such an approach lies in the incompleteness of the experimental data. In all 16 chromosomes, the degree (number of interactions per protein) was lowest for chrIII (average=16±2; Figure 3A). A similar trend was observed for betweenness (number of times that a protein acts as a bridge along the shortest path between two other proteins: average=1800±300; Figure 3B). Furthermore, chrIII is one of the chromosomes with the largest mean value for the average neighbor degree (average=380±40; Figure 3C). Our network analyses confirm chrIII as a good choice for a first line of defense against high temperature because the proteins encoded on this chromosome play less essential roles for the overall PPI network. However, once again, this portrays the duplication as a solution with least possible damage without positively suggesting causation.

Conclusions

Organisms can duplicate the whole genome or particular chromosomes (aneuploidy) in response to sudden dramatic changes in the environment. As such coarse-grained major changes are costly, aneuploidy tends to give way to more fine-tune focused solutions that require many generations to evolve. The entire chromosome III and two fragments from chromosomes IV and XII in a culture of budding yeast (*S. cerevisiae*) were duplicated as a "transient evolutionary solution" in response to high temperature - a "transition" that fostered the survival of between 400 and 2,000 generations. Here, we reported that while the proteins on all 16 main chromosomes from yeast have similar length, they differ substantially in the fraction of proteins with long regions predicted to contain protein disorder (≥30–80 consecutive residues predicted as disordered by IUPred and MD). We found the regions duplicated under heat stress depleted of predicted disorder. In fact, chromosome III was one of the two chromosomes with the least disorder (Figure 1). The other (chromosome X) is twice as large, *i.e.* would cost twice to duplicate. Decreasing the overall content in protein disorder is likely an important strategy to protect against heat stress. A detailed analysis of the experimentally characterized PPI network in yeast revealed the duplicated proteins to be connected less than average (Figure 3). The PPI analysis, therefore, added to the explanation that the duplication causes minimal damage. However, why did the duplication create an advantage under heat stress? Surprisingly, we found no sustained evidence for a significant over-representation of HSPs in the duplication i.e. of proteins that usually help out under such stress. Instead, a Gene Ontology (GO) enrichment analysis suggested that the duplicated regions were enriched in processes related to reproduction and to the import of nutrients (Figure 2). The enrichment was strongest for proteins related to sexual reproduction although the heat stress survival was maintained through budding, *i.e.* through asexual reproduction. Nevertheless, the set of GO enriched proteins appeared so important that they were duplicated although high in disorder. This might point to where the explanation for the duplication might be found. Overall, our data suggested a very simple

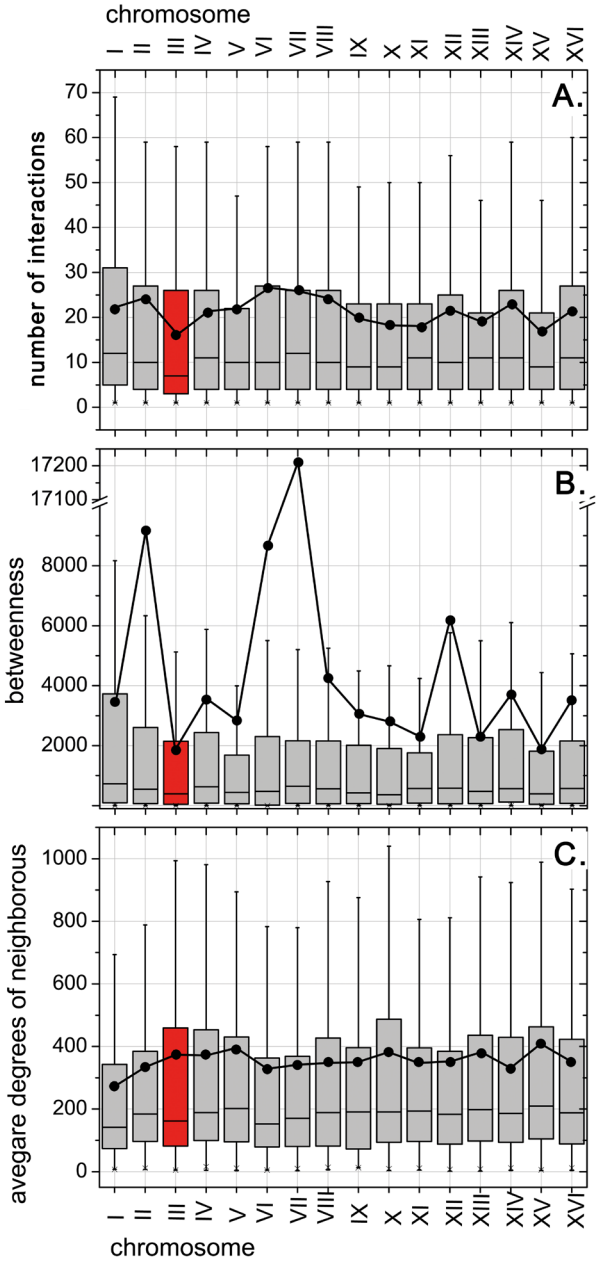

Figure 3. PPI network differs between yeast chromosomes. We began with the entire network of all PPIs with experimental annotations in yeast (Methods), and then differentially analyzed major network features: (**A**) Degree: The number of PPIs per protein (degree) was minimal for the proteins from chrIII (box in red; lowest mean - black dot and lowest median - black line in the box. (**B**) Betweenness: betweenness (number of times that a node acts as a bridge along the shortest path between two other nodes) was also lowest for chrIII. (**C**) Average neighbor degree: plotting the average degree for all network neighbors of all proteins on chrIII (i.e. all those proteins in direct PPI with proteins on chrIII), we observed a much less differentiated view. For this network feature, the proteins from chrIII had one of the highest means (black dot), but one of the lowest medians. Clearly, the proteins from the HSR-duplicated chromosome appeared less involved in the yeast network than expected by chance.

algorithm: identify the region with lowest protein disorder that is large enough, yet not too large and duplicate it along with possibly other fragments that are also depleted of disorder in order to cope with heat stress.

Author contributions

EV and BR conceived the study and designed the experiments. EV, ZG, YD, TG and MJ carried out the research. YD and TG provided expertise in protein-protein interactions and protein localization prediction respectively. EV prepared the first draft of the manuscript. EV, ZG and TG contributed to the graphics and preparation of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

This work, and all authors were supported by the German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Thanks to Tim Karl and Laszlo Kajan (TUM) for invaluable help with hardware and software; to Marlena Drabik and Inga Weise (TUM) for administrative support; to Tobias Hamp and Edda Kloppmann for helpful comments on the manuscript. A special thanks goes to the Pilpel-lab (Weizman Inst. Rehovot, Israel), in particular to Yitzhak Pilpel (Weizman) and Orna Dahan (Weizman) for the data and crucial help. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

Supplementary material

Supplementary material for Vicedo *et al.*, 2015 'Protein disorder reduced in *Saccharomyces cerevisiae* to survive heat shock'.

Table of Contents for Supplementary Material

Figure S1: Number of genes per chromosome in *Saccharomyces cerevisiae* (yeast).

Figure S2: Fragments with less disorder than proteins duplicated during heat shock response (HSR).

Figure S3: Distribution of Heat Shock proteins (HSP) along chromosomes.

Figure S4: Complete set of known GO (Gene Ontology) terms for a fragment of the HSR-duplicated chromosome IV.

Figure S5: Disordered proteins differentiate by chromosomes and roles.

Figure S6: Disorder difference between chromosome III proteins and their paralogs.

Figure S7: Distribution comparison of chromosome III proteins and the complete yeast proteome across localization classes.

Figure S8: Distribution of disordered/ordered nuclear proteins and other yeast proteins.

Table S1: General information about yeast chromosomes.

Table S2: Disorder abundance content.

Table S3: Overrepresented GO terms for molecular function analysis of chromosome III.

Table S4: Overrepresented GO terms for biological process analysis of chromosome III.

Table S5: Overrepresented GO terms for molecular function of the duplicated fragment of chromosome IV.

Table S6: Overrepresented GO terms for biological process of the duplicated fragment of chromosome IV.

Table S7: Overrepresented GO terms for biological process of the duplicated fragment of chromosome XII.

Table S8: Heat shock proteins distribution on chromosomes.

Click here to access the data.

References

1. Goffeau A, Barrell BG, Bussey H, *et al.*: **Life with 6000 genes.** *Science.* 1996; **274**(5287): 546, 563–7.
PubMed Abstract | Publisher Full Text

2. Alberghina L, Mavelli G, Drovandi G, *et al.*: **Cell growth and cell cycle in *Saccharomyces cerevisiae*: basic regulatory design and protein-protein interaction network.** *Biotechnol Adv.* 2012; **30**(1): 52–72.
PubMed Abstract | Publisher Full Text

3. Kitano H: **Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology.** *Curr Genet.* 2002; **41**(1): 1–10.
PubMed Abstract | Publisher Full Text

4. Westerhoff HV, Palsson BO: **The evolution of molecular biology into systems biology.** *Nat Biotechnol.* 2004; **22**(10): 1249–52.
PubMed Abstract | Publisher Full Text

5. Torres EM, Sokolsky T, Tucker CM, *et al.*: **Effects of aneuploidy on cellular physiology and cell division in haploid yeast.** *Science.* 2007; **317**(5840): 916–24.
PubMed Abstract | Publisher Full Text

6. Pavelka N, Rancati G, Zhu J, *et al.*: **Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast.** *Nature.* 2010; **468**(7321): 321–5.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Gordon DJ, Resio B, Pellman D: **Causes and consequences of aneuploidy in cancer.** *Nat Rev Genet.* 2012; **13**(3): 189–203.
PubMed Abstract | Publisher Full Text

8. Gresham D, Desai MM, Tucker CM, *et al.*: **The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast.** *PLoS Genet.* 2008; **4**(12): e1000303.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Hughes TR, Roberts CJ, Dai H, *et al.*: **Widespread aneuploidy revealed by DNA microarray expression profiling.** *Nat Genet.* 2000; **25**(3): 333–7.
PubMed Abstract | Publisher Full Text

10. Rancati G, Pavelka N, Fleharty B, *et al.*: **Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor.** *Cell.* 2008; **135**(5): 879–93.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Selmecki A, Gerami-Nejad M, Paulson C, *et al.*: **An isochromosome confers drug resistance *in vivo* by amplification of two genes, *ERG11* and *TAC1*.** *Mol Microbiol.* 2008; **68**(3): 624–41.
PubMed Abstract | Publisher Full Text

12. Dunham MJ, Badrane H, Ferea T, *et al.*: **Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A.* 2002; **99**(25): 16144–9.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Polakova S, Blume C, Zarate JA, *et al.*: **Formation of new chromosomes as a virulence mechanism in yeast *Candida glabrata*.** *Proc Natl Acad Sci U S A.* 2009; **106**(8): 2688–93.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Fawcett JA, Maere S, Van de Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event.** *Proc Natl Acad Sci U S A.* 2009; **106**(14): 5737–42.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Torres EM, Williams BR, Amon A: **Aneuploidy: cells losing their balance.** *Genetics.* 2008; **179**(2): 737–46.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Sheltzer JM, Blank HM, Pfau SJ, *et al.*: **Aneuploidy drives genomic instability in yeast.** *Science.* 2011; **333**(6045): 1026–30.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Torres EM, Dephoure N, Panneerselvam A, *et al.*: **Identification of aneuploidy-tolerating mutations.** *Cell.* 2010; **143**(1): 71–83.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Sheltzer JM, Amon A: **The aneuploidy paradox: costs and benefits of an incorrect karyotype.** *Trends Genet.* 2011; **27**(11): 446–53.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Pavelka N, Rancati G, Li R: **Dr Jekyll and Mr Hyde: role of aneuploidy in cellular adaptation and cancer.** *Curr Opin Cell Biol.* 2010; **22**(6): 809–15.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Yona AH, Manor YS, Herbst RH, *et al.*: **Chromosomal duplication is a transient evolutionary solution to stress.** *Proc Natl Acad Sci U S A.* 2012; **109**(51): 21010–5.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Anfinsen CB, Scheraga HA: **Experimental and theoretical aspects of protein folding.** *Adv Protein Chem.* 1975; **29**: 205–300.
PubMed Abstract | Publisher Full Text

22. Karplus M, Weaver DL: **Protein-folding dynamics.** *Nature.* 1976; **260**(5550): 404–6.
PubMed Abstract | Publisher Full Text

23. Levitt M, Chothia C: **Structural patterns in globular proteins.** *Nature.* 1976; **261**(5561): 552–8.
PubMed Abstract | Publisher Full Text

24. Levitt M, Warshel A: **Computer simulation of protein folding.** *Nature.* 1975; **253**(5494): 694–8.
PubMed Abstract | Publisher Full Text

25. Schlessinger A, Schaefer C, Vicedo E, *et al.*: **Protein disorder--a breakthrough invention of evolution?** *Curr Opin Struct Biol.* 2011; **21**(3): 412–8.
PubMed Abstract | Publisher Full Text

26. Wright PE, Dyson HJ: **Linking folding and binding.** *Curr Opin Struct Biol.* 2009; **19**(1): 31–8.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Dunker AK, Cortese MS, Romero P, *et al.*: **Flexible nets. The roles of intrinsic disorder in protein interaction networks.** *FEBS J.* 2005; **272**(20): 5129–48.
PubMed Abstract | Publisher Full Text

28. Dosztányi Z, Csizmok V, Tompa P, *et al.*: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.** *Bioinformatics.* 2005; **21**(16): 3433–4.
PubMed Abstract | Publisher Full Text

29. Schlessinger A, Liu J, Rost B: **Natively unstructured loops differ from other loops.** *PLoS Comput Biol.* 2007; **3**(7): e140.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Dunker AK, Silman I, Uversky VN, *et al.*: **Function and structure of inherently disordered proteins.** *Curr Opin Struct Biol.* 2008; **18**(6): 756–64.
PubMed Abstract | Publisher Full Text

31. Singh GP, Dash D: **Intrinsic disorder in yeast transcriptional regulatory network.** *Proteins.* 2007; **68**(3): 602–5.
PubMed Abstract | Publisher Full Text

32. Fuxreiter M, Tompa P, Simon I, *et al.*: **Malleable machines take shape in eukaryotic transcriptional regulation.** *Nat Chem Biol.* 2008; **4**(12): 728–37.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Liu J, Tan H, Rost B: **Loopy proteins appear conserved in evolution.** *J Mol Biol.* 2002; **322**(1): 53–64.
PubMed Abstract | Publisher Full Text

34. Devos D, Dokudovskaya S, Williams R, *et al.*: **Simple fold composition and modular architecture of the nuclear pore complex.** *Proc Natl Acad Sci U S A.* 2006; **103**(7): 2172–7.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Radivojac P, Iakoucheva LM, Oldfield CJ, *et al.*: **Intrinsic disorder and functional proteomics.** *Biophys J.* 2007; **92**(5): 1439–56.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Dunker AK, Gough J: **Sequences and topology: intrinsic disorder in the evolving universe of protein structure.** *Curr Opin Struct Biol.* 2011; **21**(3): 379–81.
PubMed Abstract | Publisher Full Text

37. Vicedo E, Schlessinger A, Rost B: **Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes.** *PLoS One.* 2015; **10**(8): e0133990.
PubMed Abstract | Publisher Full Text | Free Full Text

38. UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res.* 2012; **40**(Database issue): D71–5.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets.** *Nucleic Acids Res.* 2003; **31**(13): 3789–91.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Cherry JM, Hong EL, Amundsen C, *et al.*: **Saccharomyces Genome Database: the genomics resource of budding yeast.** *Nucleic Acids Res.* 2012; **40**(Database issue): D700–5.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Schlessinger A, Punta M, Rost B: **Natively unstructured regions in proteins identified from contact predictions.** *Bioinformatics.* 2007; **23**(18): 2376–84.
PubMed Abstract | Publisher Full Text

42. Schlessinger A, Punta M, Yachdav G, *et al.*: **Improved disorder prediction by combination of orthogonal approaches.** *PLoS One.* 2009; **4**(2): e4433.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Dosztanyi Z, Csizmok V, Tompa P, *et al.*: **The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.** *J Mol Biol.* 2005; **347**(4): 827–39.
PubMed Abstract | Publisher Full Text

44. Maere S, Heymans K, Kuiper M: ***BiNGO*: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics.* 2005; **21**(16): 3448–9.
PubMed Abstract | Publisher Full Text

45. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet.* 2000; **25**(1): 25–9.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–504.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Farcomeni A: **A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion.** *Stat Methods Med Res.* 2008; **17**(4): 347–88.
PubMed Abstract | Publisher Full Text

48. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Statist.* 2001; **29**(4): 1165–88.
Publisher Full Text

49. Annaluru N, Muller H, Mitchell LA, *et al.*: **Total synthesis of a functional designer eukaryotic chromosome.** *Science.* 2014; **344**(6179): 55–8.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Tompa P, Csermely P: **The role of structural disorder in the function of RNA and protein chaperones.** *FASEB J.* 2004; **18**(11): 1169–75.
PubMed Abstract | Publisher Full Text

51. Tarca AL, Bhatti G, Romero R: **A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity.** *PLoS One.* 2013; **8**(11): e79217.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Boles E, Hollenberg CP: **The molecular genetics of hexose transport in yeasts.** *FEMS Microbiol Rev.* 1997; **21**(1): 85–111.
PubMed Abstract | Publisher Full Text

53. Ozcan S, Dover J, Johnston M: **Glucose sensing and signaling by two glucose receptors in the yeast *Saccharomyces cerevisiae*.** *EMBO J.* 1998; **17**(9): 2566–73.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Reifenberger E, Freidel K, Ciriacy M: **Identification of novel *HXT* genes in *Saccharomyces cerevisiae* reveals the impact of individual hexose transporters on glycolytic flux.** *Mol Microbiol.* 1995; **16**(1): 157–67.
PubMed Abstract | Publisher Full Text

55. Brown CJ, Todd KM, Rosenzweig RF: **Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment.** *Mol Biol Evol.* 1998; **15**(8): 931–42.
PubMed Abstract | Publisher Full Text

56. Tannenbaum E: **A comparison of sexual and asexual replication strategies in a simplified model based on the yeast life cycle.** *Theory Biosci.* 2008; **127**(4): 323–33.
PubMed Abstract | Publisher Full Text

57. Zeyl C, Curtin C, Karnap K, *et al.*: **Antagonism between sexual and natural selection in experimental populations of *Saccharomyces cerevisiae*.** *Evolution.* 2005; **59**(10): 2109–15.
PubMed Abstract | Publisher Full Text

58. Nelson B, Parsons AB, Evangelista M, *et al.*: **Fus1p interacts with components of the Hog1p mitogen-activated protein kinase and Cdc42p morphogenesis signaling pathways to control cell fusion during yeast mating.** *Genetics.* 2004; **166**(1): 67–77.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Zhang M, Bennett D, Erdman SE: **Maintenance of mating cell integrity requires the adhesin Fig2p.** *Eukaryot Cell.* 2002; **1**(5): 811–22.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Erdman S, Lin L, Malczynski M, *et al.*: **Pheromone-regulated genes required for yeast mating differentiation.** *J Cell Biol.* 1998; **140**(3): 461–83.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Ramezani-Rad M: **The role of adaptor protein Ste50-dependent regulation of the MAPKKK Ste11 in multiple signalling pathways of yeast.** *Curr Genet.* 2003; **43**(3): 161–70.
PubMed Abstract | Publisher Full Text

62. Truckses DM, Bloomekatz JE, Thorner J: **The RA domain of Ste50 adaptor protein is required for delivery of Ste11 to the plasma membrane in the filamentous growth signaling pathway of the yeast *Saccharomyces cerevisiae*.** *Mol Cell Biol.* 2006; **26**(3): 912–28.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Sandler H, Kreth J, Timmers HT, *et al.*: **Not1 mediates recruitment of the deadenylase Caf1 to mRNAs targeted for degradation by tristetraprolin.** *Nucleic Acids Res.* 2011; **39**(10): 4373–86.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Berlin V, Styles CA, Fink GR: **BIK1, a protein required for microtubule function during mating and mitosis in *Saccharomyces cerevisiae*, colocalizes with tubulin.** *J Cell Biol.* 1990; **111**(6 Pt 1): 2573–86.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Miller RK, Rose MD: **Kar9p is a novel cortical protein required for cytoplasmic microtubule orientation in yeast.** *J Cell Biol.* 1998; **140**(2): 377–90.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Lahav R, Gammie A, Tavazoie S, *et al.*: **Role of transcription factor Kar4 in regulating downstream events in the *Saccharomyces cerevisiae* pheromone response pathway.** *Mol Cell Biol.* 2007; **27**(3): 818–29.
PubMed Abstract | Publisher Full Text | Free Full Text

67. Field CM, Kellogg D: **Septins: cytoskeletal polymers or signalling GTPases?** *Trends Cell Biol.* 1999; **9**(10): 387–94.
PubMed Abstract | Publisher Full Text

68. Cid VJ, Adamíková L, Cenamor R, *et al.*: **Cell integrity and morphogenesis in a budding yeast septin mutant.** *Microbiology.* 1998; **144**(Pt 12): 3463–74.
PubMed Abstract | Publisher Full Text

69. Longtine MS, DeMarini DJ, Valencik ML, *et al.*: **The septins: roles in cytokinesis and other processes.** *Curr Opin Cell Biol.* 1996; **8**(1): 106–19.
PubMed Abstract | Publisher Full Text

70. Madden K, Snyder M: **Cell polarity and morphogenesis in budding yeast.** *Annu Rev Microbiol.* 1998; **52**: 687–744.
PubMed Abstract | Publisher Full Text

71. Carroll CW, Altman R, Schieltz D, *et al.*: **The septins are required for the mitosis-specific activation of the Gin4 kinase.** *J Cell Biol.* 1998; **143**(3): 709–17.
PubMed Abstract | Publisher Full Text | Free Full Text

72. Barral Y, Parra M, Bidlingmaier S, *et al.*: **Nim1-related kinases coordinate cell cycle progression with the organization of the peripheral cytoskeleton in yeast.** *Genes Dev.* 1999; **13**(2): 176–87.
PubMed Abstract | Publisher Full Text | Free Full Text

73. Park HO, Sanson A, Herskowitz I: **Localization of Bud2p, a GTPase-activating protein necessary for programming cell polarity in yeast to the presumptive bud site.** *Genes Dev.* 1999; **13**(15): 1912–7.
PubMed Abstract | Free Full Text

74. de Oliveira IM, Henriques JA, Bonatto D: ***In silico* identification of a new group of specific bacterial and fungal nitroreductases-like proteins.** *Biochem Biophys Res Commun.* 2007; **355**(4): 919–25.
PubMed Abstract | Publisher Full Text

75. Forestier C, Frangne N, Eggmann T, *et al.*: **Differential sensitivity of plant and yeast MRP (ABCC)-mediated organic anion transport processes towards sulfonylureas.** *FEBS Lett.* 2003; **554**(1–2): 23–9.
PubMed Abstract | Publisher Full Text

76. Chen PW, Fonseca LL, Hannun YA, *et al.*: **Coordination of rapid sphingolipid responses to heat stress in yeast.** *PLoS Comput Biol.* 2013; **9**(5): e1003078.
PubMed Abstract | Publisher Full Text | Free Full Text

77. Peifer C, Sharma S, Watzinger P, *et al.*: **Yeast Rrp8p, a novel methyltransferase responsible for m1A 645 base modification of 25S rRNA.** *Nucleic Acids Res.* 2013; **41**(2): 1151–63.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Alepuz PM, Matheos D, Cunningham KW, *et al.*: **The *Saccharomyces cerevisiae* RanGTP-binding protein msn5p is involved in different signal transduction pathways.** *Genetics.* 1999; **153**(3): 1219–31.
PubMed Abstract | Free Full Text

79. Boyle EI, Weng S, Gollub J, *et al.*: **GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics.* 2004; **20**(18): 3710–5.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Goldberg T, Hecht M, Hamp T, *et al.*: **LocTree3 prediction of localization.** *Nucleic Acids Res.* 2014; **42**(Web Server issue): W350–5.
PubMed Abstract | Publisher Full Text | Free Full Text

81. Ramilowski JA, Goldberg T, Harshbarger J, *et al.*: **A draft network of ligand-receptor-mediated multicellular signalling in human.** *Nat Commun.* 2015; **6**: 7866.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✓ ? ✓

Version 1

Referee Report 20 November 2015

doi:10.5256/f1000research.7734.r11125

Anuj Kumar
Department of Molecular, Cellular & Developmental Biology, University of Michigan, Ann Arbor, MI, USA

The manuscript by Vicedo and colleagues presents an interesting observation: the authors have examined chromosomal regions (all of chromosome III and fragments of chromosomes IV and XII) that are duplicated in *Saccharomyces cerevisiae* in response to sudden exposure to high temperature and find that these chromosomal sequences are significantly decreased for genes encoding proteins with long disordered regions. The authors further analyzed these duplicated regions and the encompassed genes for any enrichment in annotated GO terms, as well as for encoded protein positioning in interaction networks. The results do not indicate significant GO term enrichment and reveal that the encoded proteins exhibit a decreased number of interactions per protein. The biological advantage to this duplication remains unclear.

Comments/suggestions:

The main conclusion presented here is interesting, but as the authors themselves attest, this observation does not explain a biological advantage behind the duplication.

On p. 4, the authors state that introducing an extra copy of *HSP30* into wild-type yeast does not modify the ability of the cells to cope with high temperature. The inclusion of laboratory data considering the effect of adding an extra copy of genes or chromosomal regions corresponding to some of the duplicated sequences would strengthen the paper significantly. This seems to be the easiest way to address a biological effect from duplication of a given gene.

In regards to the analysis, are the observed GO function annotations enriched with respect to other chromosomes/segments as opposed to being enriched against the genome as a whole? If the advantage to the cell centered on the functions associated with the genes in the duplicated regions, then these regions relative to other regions may be enriched for a function. If I'm thinking of this correctly, that would be slightly different than comparing a region for enrichment against the whole genome. Maybe the authors could compare enrichment in one chromosome versus another or utilize a sliding window corresponding to the size of a duplicated fragment to identify regions that would be most enriched for some potentially interesting functions. That might be a more sensitive means of identifying a functional enrichment for the duplicated regions.

Typos/stylistic suggestions:

- on p.2, first line under Introduction: I think it would be sufficient to state “The baker’s yeast *Saccharomyces cerevisiae*” rather than the text in parentheses.
- on p. 3, fourth paragraph under “Duplications in response to high temperature reduce protein disorder”: the first sentence in this paragraph (“Assume a certain amount …”) needs to be reworded.
- on p. 4, first paragraph, line 16: delete “the” from “insignificant the finding”

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 17 November 2015

doi:10.5256/f1000research.7734.r11121

Paul Pavlidis
Centre for High-Throughput Biology and Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada

Vicedo *et al*. report a computational analysis sparked by the interesting findings of Yona *et al.* (2012) of yeast duplication of chrIII having a selective advantage in the face of heat stress. Yona *et al*. did not fully mechanistically explain the reason chrIII aneuploidy is the one selected for, so Vicedo *et al*. have proposed a hypothesis: that chrIII has a substantially lower number of disordered proteins. They test this computationally, followed by some additional bioinformatics and “by hand” characterization of chrIII genes (and some other regions of interest following from Yona *et al*.).

The difficulty here is assigning cause vs. “permissive”. As Vicedo *et al*. report, the disorder hypothesis has limited predictive value because chrX genes also have a low disorder (on average), so the size of the chromosome is posed as the other important variable. However, Vicedo *et al*. seem to be proposing that “low disorder” is good for heat resistance per se (I grant them this) – and that overexpression of low disorder proteins is even better. I have difficulty with this second step, because the way the experiment of Yona *et al*. was done, it could easily be that there are “heat resistance proteins” on chrIII and that the overall duplication of chrIII is tolerated in the context of the advantage of overexpression those genes. But if this was the end of the story it would be hard to make a determination of whether this is a viable hypothesis.

However, there is an obvious problem: the work of Yona *et al.* identified 17 genes on chrIII that appear to be the main culprits for the heat resistance (at least most of them). I see no mention of these 17 in Vicedo *et al*. nor of the 22 control genes tested by Yona *et al*. If Vicedo *et al.* are right then there should be a difference in the disorder of these two sets of proteins. Otherwise, the observations might still be relevant, but that the orderedness of chrIII proteins might be permissive for overexpression of the actual heat-resistance genes via aneuploidy. In that case it might be the rest of the proteins on chrIII that have the orderedness properties, not the 17. (Note that I was not familiar with the Yona work before this review and I have not checked to see if Yona *et al.* or others have done any followup.)

Given the omission of discussion of the 17, the sections of this paper on network analysis, GO and

localization cannot be interpreted with confidence. While I have some quibbles about them I would rather wait to see the response to the comments above.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 17 November 2015

doi:10.5256/f1000research.7734.r11122

Melchor Sanchez-Martinez
Mind the Byte, Barcelona, Spain

The research article entitled 'Protein disorder reduced in *Saccharomyces cerevisiae* to survive heat shock' by Vicedo, Rost and co-workers shows how *Saccharomyces cerevisiae* reduces protein disorder to survive heat shock. It constitutes an interesting example about the usage of bioinformatic techniques to analyze protein disorder and its implications at a whole proteome level. In the article, there is a comprehensive explanation of study design, methods and analysis. The conclusions are well explained and justified on the basis of the results.

Consequently, the manuscript is recommended for approval. It is a good piece of science that meets the indexation requirements of F1000Research.

However I have some comments that the authors may consider and/or answer.

1. As far as I know except some rare exceptions the protein disorder is reduced as temperature increases, oppositely as happens with ordered proteins or protein regions. With increasing temperature, disordered proteins and regions tend to adopt a transitory structure. Commonly this transitory structure is necessary for proteins to perform its biological function. In other recent works that the authors have published have published (Reference 37 in the References section), they stated that "protein disorder appeared as a possible building block to bring about evolutionary changes such as the adaptation to different habitats" and in that sense seems that more disorder should imply a better response to heat shock.

   Thus is surprising for me that in response to heat there is a protein disorder reduction, whereas I expect a disorder increment. Why does it happens? Maybe the answer is so easy as that the disordered proteins do not help to "fight" against heat shock or as the authors said "...Some of these proteins with long disordered regions might not work correctly in heat...", but I am curious about that. Do you have any evidence or supported hypothesis to explain that?

2. Regarding to authors statement "...Some of these proteins with long disordered regions might not work correctly in heat...", a plausible way to study that and obtain a more conclusive answer could be to perform a molecular simulation. Maybe a Replica Exchange Molecular Dynamics or Monte Carlo simulations could give a better understanding of what happens with these protein at high temperatures.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.