Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

PROTEIN DOMAIN PHYLOGENIES
Information Theory and Evolutionary Dynamics

K. Hamacher
Bioinformatics & Theoretical Biology Group, Technische Universität Darmstadt, Germany
hamacher@bio.tu-darmstadt.de

Keywords: information theory; Jensen-Shannon divergence; evolutionary dynamics; phylogenetic trees; SUPFAM

Abstract: The ever-increasing wealth of whole-genome information prompts for phylogenies based on entire genomes. The quest for a good distance measure, however, poses a big challenge; e.g. because of large-scale evolutionary events such as genomic rearrangements or inversions. We introduce here an information theory driven measure that for the encoded protein domain composition of genomes as protein domains are key evolutionary entities. Thus the new method focuses on *selective advantageous* events. As evolving different protein domain compositions is more complex than single point mutations, the method makes longer evolutionary times accessible. Illustrating the new methodology we extract several phylogenetic trees for some 700 genomes, e.g. the separation of the three kingdoms of life, trees for mammals and bacillales, and a speculative result for plants (monocotyledons and dicotyledons). The method itself is shown to be robust against incomplete genome sampling. It has a consistent interpretation in both, information space at the sequence/information level and at the level of stochastic, evolutionary dynamics. In contrast to established protocols it becomes more accurate as more organisms are taken into account. Finally we show the equivalence to a (simplified) model of evolutionary dynamics of proteomes.

1 INTRODUCTION

In the post-genome era the availability of vast amounts of genetic, proteomic, metabolic, and biochemical data opens new horizons for investigation into the origin of life and its evolution. To this end a detailed understanding of phylogenetic relationships is mandatory. These phylogenies provide further insight into the evolutionary trait of species and form to a large extent our picture of the biological history of life (Woese, 2000; Woese, 2002).

Now the availability of whole genome data provides for the opportunity of better phylogenetic insight (Philippe et al., 2005) as recently discussed e.g. by Yang et al. (Yang et al., 2005) For example Tekaia et al. (Tekaia et al., 1999) used whole genomes for phylogenetic analysis. Their suggested method deals, however, not with the proteome (see below for advantages) and compares general gene products based on sequences. Other approaches based on whole-genomes are the studies by Otu and Sayood using Lempel-Zif-compression-complexity of the genomic sequences (Otu and Sayood, 2003), the related idea by Li et al. (Li et al., 2001; Li et al., 2004) based on Kolmogorov complexity, and the subsequent refinement by Mantaci and co-workers (Mantaci et al., 2008). All the approaches are sequence-focused, thus will have difficulties to account for e.g. constraints on the gene products in the physical realm e.g. protein biophysics, structural biological issues and so on. In addition although the concept of Kolmogorov complexity is well established in theoretical computer science and has proven to be of great value there, it is not possible to compute its value numerically, but one has instead to approximate it (Li and Vitányi, 1997).

1.1 Aligned-Sequence vs. Whole-Genome Phylogenies

In general the signal-to-noise-ratio of methods based on proteomic characteristics will be better than for e.g. the well-established 16S-RNA-phylogenies. This

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

effect is due to the stable conservation of protein(-domain) structure in comparison to its coding sequence. Related to this is the additional advantage that longer evolutionary time-scales will become accessible. As noted by Yang et al. (Yang et al., 2005) "house-keeping" proteins such as metabolic enzymes, cytoskeleton-proteins, or histones can be expected to evolve even more slowly and thus increase the signal-to-noise-ratio further. From a conceptional point of view we have to concede that evolution during selection acts upon advantageous changes of the phenotype and not upon changes in the genotype. The phenotype is, however, to a large extent determined by the encoded proteome. Sequence based phylogenies are focused solely on mutational events and their fixation; proteome-based approaches are more concerned with the (realized) selective advantage.

1.2 Difficulties due to Sequence Alignments in Traditional Phylogenetic Algorithms

In typical sequence based phylogenetic methods accurate sequence alignments pose the biggest challenge due to alignment sensitivity towards parameter changes in the alignment procedure. This issue was extensively discussed e.g. by Li et al. (Li et al., 2001). In particular the assumption on a universal applicability of substitution matrices for longer time scales was disputed.

By looking at proteomic information we - at least theoretically - avoid this problem completely. We will in particular map the genome to protein domain composition (PDC) vectors that quantify the abundance of protein domain folds coded within the species' genomes. As will laid out below we rely on Hidden-Markov-Models (HMMs) to extract proteomes from genomic sequences and thus work only implicitly with those alignments used to derive the HMMs. These alignments for HMMs are, however, based on a broader sample than we will face in phylogenetic analysis and will thus be able to reduce noise.

1.3 Sequence-Based Whole-Genome Approaches

Gerstein (Gerstein, 1998) pioneered the usage of proteomic information for phylogenies with his seminal work on the eight genomes available at that time. His distance measure is based just on the number of shared protein folds. While this avoids the problem of unknown protein folds or domains that one might fail to identify in the genome, it will on the other hand fail to discriminate between closely related organisms as they can be expected to command over the same collection of protein folds, but with varying abundances. In particular the argument about unknown protein folds becomes less compelling over time as the known, structural space of proteins has become denser over the last years. Compare for example to the hypothesis that the accessible protein fold space is completely known (Zhang and Skolnick, 2005): based on this assumption, one can argue that the argument about unknown protein domains is not relevant nowadays.

Yang et al. (Yang et al., 2005) used as a distance measure the number of shared superfamily folds. This ansatz constitutes a yes/no-decision for the individual fold and neglects abundances of folds. A binary decision such as this is sensitive with respect to the accuracy of protein prediction. In addition - as with Gerstein's initial idea - it will in general not be able to grasp subtle differences between closely related species.

Additional studies were undertaken by Fukami-Kobyashi et al. (Fukami-Kobayashi et al., 2007) and by Fong et al. (Fong et al., 2007). The latter work bases its distance measure on genomic rearrangements represented by networks of necessary domain rearrangements. The issue of elevated rates of rearrangements for different organisms (Ekman et al., 2007) might, however, restrict the applicability of such a method to narrowed sets of taxa.

1.4 Proteome Information as a Distance Measure

In the following we will derive phylogenetic distances from the composition of the proteomes of organisms as represented by the protein domain composition (PDC) vectors introduced above. To this end we compute the relative frequencies p_i of protein domain folds i within the proteome of an organism. For two such distributions (p and q) for two organisms we can use the Jensen-Shannon entropy $H_{\mathrm{JS}}(p,q)$ as an information theoretical distance measure. H_{JS} reads (Lin, 1991):

$$H_{\mathrm{JS}}(p,q) \quad := \quad \frac{1}{2}\cdot D_{\mathrm{KL}}(p\,|\,m)+\frac{1}{2}\cdot D_{\mathrm{KL}}(q\,|\,m) \quad (1)$$

Here D_{KL} is the Kullback-Leibler-divergence (see Methods section for details) and $m := \frac{1}{2}\cdot p+\frac{1}{2}\cdot q$ is an average distribution, which can be interpreted as an ancestor at an evolutionary branching point.

The Jensen-Shannon entropy gives the evolutionary distance as the amount of information that one most provide to describe the difference from PDC p

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

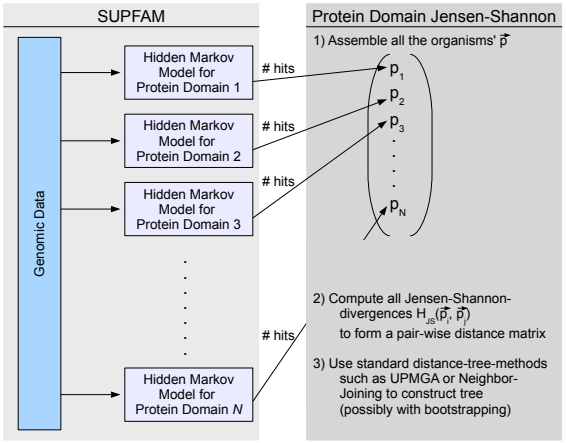

Figure 1: The mapping of a collection of genomes to a phylogenetic tree. Starting from genomic data we derive the number of contained protein domains by the Hidden Markov Models of the SUPFAM database. These values give rise to a protein domain composition vector $\vec{p}$ for each species. The species are then compared by pair-wise distance computations via the Jensen-Shannon-divergence $H_{\rm JS}(\vec{p}_i, \vec{p}_j)$ of the respective protein domain composition vectors $\vec{p}_i$.

to the PDC q through an intermediate, 'common ancestor' of average PDC m. The entropy $H_{\rm JS}$ is usually measured in bits and its statistical properties for artificially created sequences was discussed in (Grosse et al., 2002). Figure 1 gives an overview of the protocol.

In the Methods section we discuss the derivation of $H_{\rm JS}$ and in the Discussion section we show how this measure can be rationalized by a simple model of evolutionary dynamics in the proteome/protein domain fold space. Derivation of trees from $H_{\rm JS}$-distances can be achieved with the well-known PHYLIP-package (Felsenstein, 1989).

2 METHODS

2.1 Genome to Proteome conversion

Hidden Markov Models (HMMs) form the basis of the Superfamily database (SUPFAM (Gough et al., 2001; Wilson et al., 2007)). These HMMs search the NCBI Entrez Genome database and identify protein superfamily folds. These folds are based on the Structural Classification of Proteins database (SCOP) reflecting the hierarchy of protein domains in the PDB. The superfamily level indicates common ancestry.

Each fold identification by the SUPFAM-HMMs is assigned a measure of the reliability. This quantity is given as the well-known E-value of confidence. We applied a threshold on E-values of $E \leq 10^{-4}$ for the identification of protein domains by the HMMs. In accordance with a previous study (Yang et al., 2005) we found a difference in the distances when based on all HMM results or just the ones below the threshold (see supporting figure S1). As was argued in (Yang et al., 2005) the choice of $E \leq 10^{-4}$ increases the accuracy and the robustness of the genome-to-proteome-mapping. The flow of information is shown in Figure 1.

2.2 Computation of distances

The well-known Kullback-Leibler information divergence (MacKay, 2004) measures the relative information of a probability distribution q with respect to a reference distribution p - both defined on the same event set $\mathcal{X}$.

The Kullback-Leibler-divergence is a universal measure with a wide range of potential applications, from information driven sequence analysis (Lund et al., 2005) to ligand design (Hamacher, 2007c; Hamacher et al., 2006). Benos, Bulyk, and Stormo used the Kullback-Leibler-divergence in a validation study on protein-DNA-interaction and discussed its relation for closely related distributions to the χ^2-distribution (Panayiotis V. Benos, Alan S. Lapedes and Gray D. Stormo, 2002) - eventually proving that the Kullback-Leibler-divergence is extensive. The latter property is most relevant as we can immediately conclude that in general finite-size-effects will not play any important role. Burstein et al. in fact used the Kullback-Leibler-divergence as a distance measure for *sequences* of whole genomes (Burstein et al., 2005).

The Jensen-Shannon entropy was introduced in a seminal paper by Lin (Lin, 1991) and reads in general for a parameter $\lambda \in [0;1]$:

$$H_{\rm JS}(p,q) := \lambda \cdot D_{\rm KL}(p \,|\, \lambda \cdot p + (1-\lambda)\cdot q) + (1-\lambda)\cdot D_{\rm KL}(q \,|\, \lambda \cdot p + (1-\lambda)\cdot q) \quad (2)$$

Here $D_{\rm KL}$ is the Kullback-Leibler-divergence (MacKay, 2004) and reads $D_{\rm KL}(p \,|\, m) := \sum_i p_i \cdot \log_2 \frac{p_i}{m_i}$. $D_{\rm KL}$ is a relative entropy and $H_{\rm JS}$ therefore a linear combination of such entropies. $D_{\rm KL}$ itself would not be a suitable distance measure as it is not symmetric under exchange of its arguments, as in general $D_{\rm KL}(p \,|\, m) \neq D_{\rm KL}(m \,|\, p)$.

For $\lambda = 1/2$ we have, however, $H_{\rm JS}(p,q) = H_{\rm JS}(q,p)$ and thus the Jensen-Shannon entropy $H_{\rm JS}(p,q)$ is for this particular λ-value a symmetrized version of the Kullback-Leibler-divergence.

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

In addition the properties

- $D_{\mathrm{KL}}(p,q) \geq 0$ and
- $D_{\mathrm{KL}}(p,q) = D_{\mathrm{KL}}(q,p)$ if and only if the probabilities p and q are on their domain of definition equal with probability one ($\forall_i\ p_i = q_i$).

are inherited by $H_{\mathrm{JS}}(p,q)$. Note that H_{JS} is not a metric (it does not fulfill the triangle inequality), but $\sqrt{H_{\mathrm{JS}}}$ is (Endres and Schindelin, 2003). Although one would therefore tend to use $\sqrt{H_{\mathrm{JS}}}$, we work with H_{JS} as it has a direct interpretation in the framework of stochastic processes and we can therefore give a direct interpretation of the obtained H_{JS}-values.

2.3 Construction of the phylogenetic tree based on the computed distance matrix

The phylogenetic trees were built from the above described distances using the Neighbor-Joining-Method as implemented in the neighbor-program of the PHYLIP-package (Felsenstein, 1989) which is the last step of the protocol of figure 1.

Furthermore we performed bootstrapping (Soltis and Soltis, 2003) on the obtained data using a noise term of relative change of 5% in the distance values to mimic potential inaccuracies of the SUPFAM Hidden Markov Models or non-complete coverage of the used genomes or both. Consensus trees were then again determined with the consensus-program of the PHYLIP-software and are shown in figures 2 to 6.

3 RESULTS

3.1 Obtained phylogenetic relations

As the first fundamental application of phylogenetic methods we looked at the kingdoms of life (archae, bacteria, eukaryota). We found a bootstrap support of 100% for the separation of the three domains of life (Archaea, Bacteria, Eukaryotes). We show this and the overall emerging picture in figure 2.

We furthermore extracted subtrees for mammals, insects, and bacillales to illustrate the correctness and precision of the proposed method. The results are shown in figures 3, 4, and 5. These trees are in very good agreement with previously derived classifications, but partially show higher support on some clades.

Here we want to comment briefly on some additional aspects observed during these investigations:

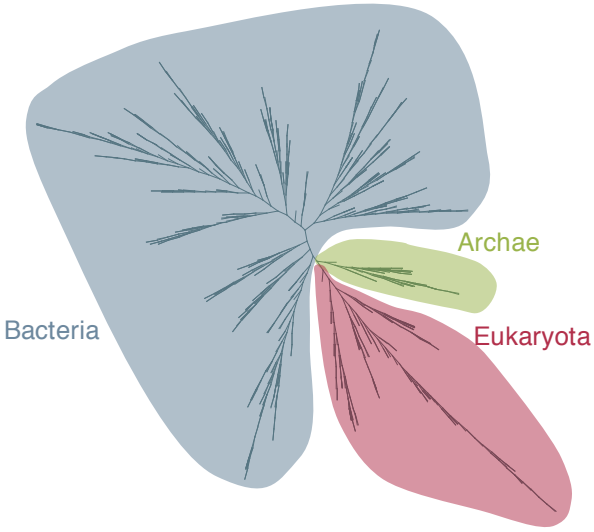

Figure 2: The separation of the three kingdoms of life obtained with 100% bootstrap support at the major branching points for all 698 species.

- The mammalian consensus tree is accurate, effectively grouping *H. sapiens* and *P. troglodytes*, putting *M. mulatta* (Rhesus monkey) in their neighborhood. Also the pairing of *R. norvegicus* and *M. musculus* is reasonable, suggesting a subtree of rodents if more genomic data were available. We note in passing that the mammalian tree did not show much support when we applied Yang et al.'s procedure on the noisy data described in the Methods section of this paper. This finding suggests that our distance measure is somewhat more stable in regard to incomplete, noisy mapping of genomes to protein domain fold space / PDCs.
- For insects the grouping of the two strains of *D. melanogaster* and the phylogenetic vicinity of *D. psuedoobscura* indicate the consistency of the methodology presented here. In addition putting *A. aegypti* (yellow fever mosquito) and *C. pipiens quinquefasciatus* (southern house mosquito) with a perfect support of 100% together is evident

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

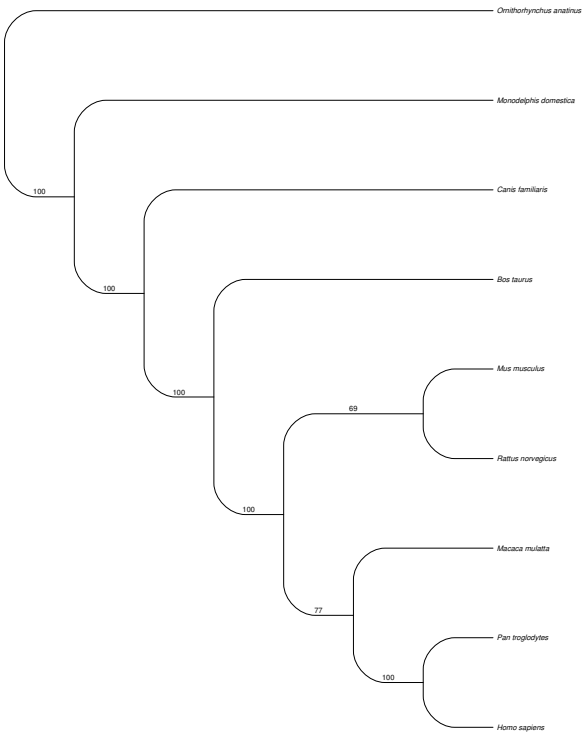

Figure 3: Phylogenetic tree for mammalian genomes. The numbers at the arcs indicate the percentage bootstrap support (with noise) for the clades.

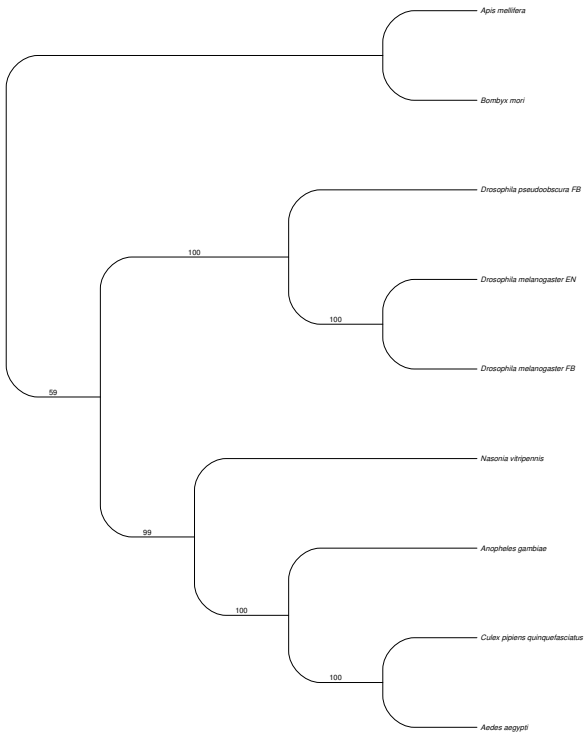

Figure 4: Phylogenetic tree for insects again with percentage bootstrap support for the clades. Abbreviations: FB - FlyBase, EN - Ensmbl.

due to their close relationship; this is furthermore augmented by the nearby *A. gambiae* (morphologically indistinguishable mosquitoes).

- For bacillales the methodology presented here is able to sub-divide the bacillales correctly. As a first indication we observe that the 12 *Staphylococcus aureus* strains were correctly put into a common subtree. Not surprisingly the support in this subtree at the various branching points is rather low (some $50-60\%$). The support to separate this subtree from *Staphylococcus epidermidis* and *Staphylococcus haemolyticus* is in contrast again at the perfect 100% level. Other strains, such as *Bacillus licheniformis* and *Listeria monocytogenes* were also put into shared subtrees. Some support is rather low in the clade of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*. This, again, comes as no surprise as these three organisms are essentially the same, differing only in their plasmids - which were not taken into account in this study and therefore could not support separation in clades.

A somewhat more complicated picture emerges for plants. First we have to acknowledge that - with respect to the overall number of plants - there are only a few plant genomes available. This turned out to be the largest influence on the tree of plants. In figure 6 we show the resulting tree. We found *Arabidopsis thaliana* to be key to further discrimination between monocotyledons and dicotyledons. During the work the genome of another monocotyledon, *Soghorum bicolor*, became available in SUPFAM and we enriched the data set with the PDC vector of this organism, repeated the distance computation, and the derivation of the phylogenetic tree. The result is shown in the lower part of figure 6 and indicates a genetic network between monocotyledon and dicotyledons, while the previously somewhat unclear position of *A. thaliana* is clarified. The placement is resolved by additional genomes of related organisms. This supports the assertion that the method in general will profit from larger data sets including closely related organisms.

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

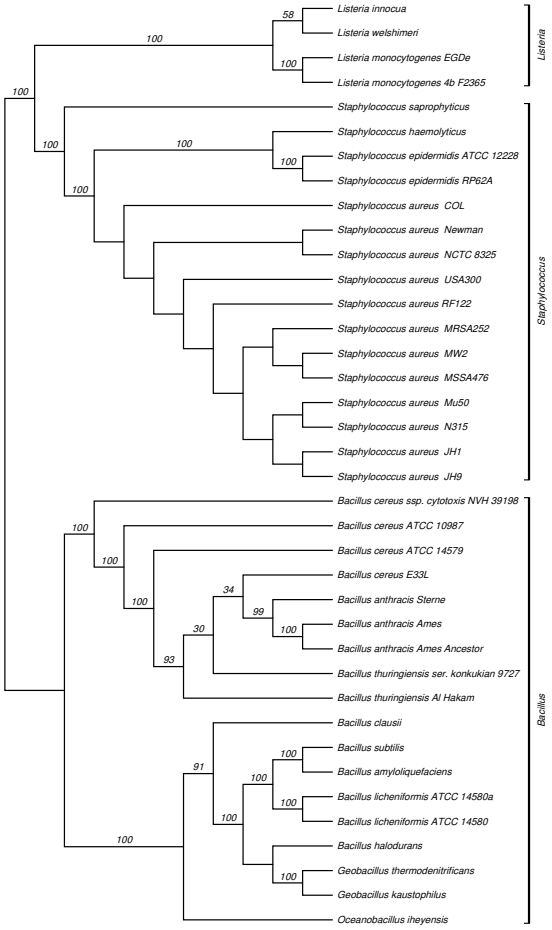

Figure 5: Phylogenetic tree for Bacillales. Again numbers represent the percentage bootstrap support. Note that *B. cereus*, *B. thuringiensis*, and *B. anthracis* differ only in their plasmids, which were not taken into account in the computation of the phylogenetic distance.

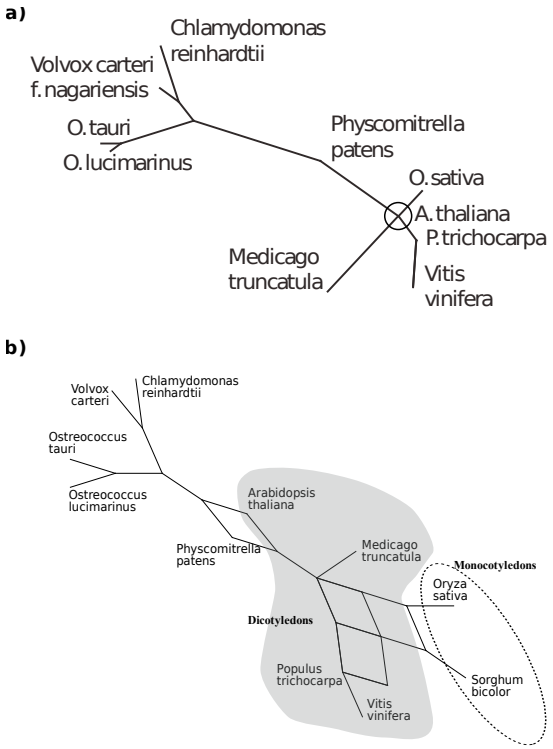

Figure 6: a) a phylogenetic tree for plants as derived by the neighboring-joining method. b) Phylogenetic Network for plants enriched by the newly available genomic data of *Sorghum bicolor*, obtained with the splits-decomposition procedure as implemented in SplitsTree (Huson and Bryant, 2006). The phylogenetic relations are much more accurately determined for the enriched data set which includes two monocotyledons. This effect is discussed in the text.

4 DISCUSSION

We have exemplified - by the plants example - a property of the presented method that is at first glance counterintuitive: the derivation of reasonable tree is simplified by a high number of closely related genomes. Usually one would assume separating closely related organisms based on just small differences in the phylogenetic distance is difficult and puts a lot of burden on the tree building algorithm. On the other hand (and this is special to our methodology) the denser the 'organism space' is populated, the more genomes are available, the more reasonable the setting of the intermediate ancestor probability distribution m in equation 1 is. There are indications that this property is also found in advanced sequence based approaches (Dunn et al., 2008).

In figure 7 we show distances obtained by our method and by the one of Yang et al. (Yang et al., 2005) Obviously there exists a correlation, which implies general agreement. Subtle differences are, however, also present and give rise to a different distribution of those values. In figure 8 we show the probability distribution of those distance measures for the 698 organisms under consideration in this study. The results of figure 8 point to a richer structure of the distance values within the all-or-nothing method by Yang et al. A richer structure implies, however, an abundance of local optima in the tree generation problem that the mapping distance matrix $\longrightarrow$ tree poses (it eventually constitutes an optimization problem). Local optima can lead to problems and suboptimal trees when applying greedy algorithms such as neighbor joining. Therefore the data in figure 8 suggests

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

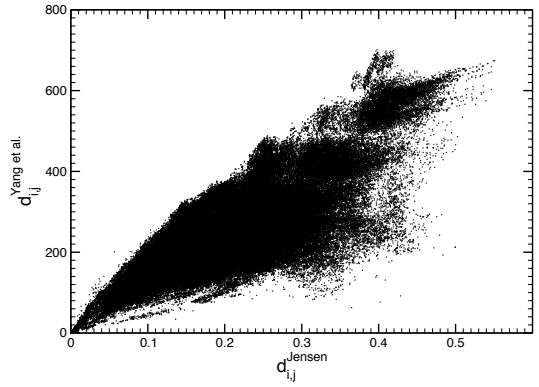

Figure 7: Scatter Plot of the Jensen Information Distances with a threshold and the absolute measure of ref. (Yang et al., 2005) for all 698 species.

a (partial) resolution by our method of the local optima problem. We expect greedy algorithms to more reliable applicable on our distance matrices.

In addition perturbations due to errors in the genome-to-PDC mapping will in general be less relevant if the distribution of distances has fewer features, because this implies - again - less bias towards local optima.

4.1 Relation to a Dynamical Model of Evolution

It can be further shown that the Jensen-Shannon entropies are equivalent to evolutionary distances in a simplistic model of protein domain evolution. To this end we define: let P_i and Q_i be the number of superfamily folds i in two organisms to be compared. Then $P = \sum_i P_i$ and $Q = \sum_i Q_i$ is the overall number of folds within the proteome of the organisms. Then $M_i = 1/2 \cdot P_i + 1/2 \cdot Q_i$ and $M = \sum_i M_i = 1/2 \cdot P + 1/2 \cdot Q$ are the value for the average, common 'ancestor'. As outlined in the Methods section we compute the relative frequencies of occurrence $p_i = P_i/P$, $q_i = Q_i/Q$, and $m_i = M_i/M$.

Then we can rewrite the Jensen-Shannon entropy as

$$\begin{aligned} H_{\mathrm{JS}}(p,q) &= \frac{1}{2}\cdot\sum_i p_i \cdot \log\frac{p_i}{m_i} + \frac{1}{2}\cdot\sum_i q_i \cdot \log\frac{q_i}{m_i} \\ &= \frac{1}{2}\cdot\langle\log\frac{p_i}{m_i}\rangle_p + \frac{1}{2}\cdot\langle\log\frac{q_i}{m_i}\rangle_q \\ &= \frac{1}{2}\cdot\Big(\langle\log\frac{P_i}{M_i}\rangle_p - \langle\log\frac{P}{M}\rangle_p \\ &\quad +\langle\log\frac{Q_i}{M_i}\rangle_q - \langle\log\frac{Q}{M}\rangle_q\Big) \\ &= \frac{1}{2}\cdot\Big(\langle\log\frac{P_i}{M_i}\rangle_p - \log\frac{P}{M} + \\ &\quad \langle\log\frac{Q_i}{M_i}\rangle_q - \log\frac{Q}{M}\Big) \end{aligned} \tag{3}$$

where $\langle\ldots\rangle_{p,q}$ indicates the expectation value of the argument with respect to the probability distributions p and q, respectively. If we now assume simplisticly that the underlying evolutionary process of domain losses and gains is of Poisson character, then Fukami-Kobayashi et al. (Fukami-Kobayashi et al., 2007) have argued that ratio like $\log\frac{P}{M}$ is the evolutionary time/distance it took to get from the PDC m to p. Similarly the fold specific terms $\log\frac{P_i}{M_i}$ are the evolutionary times it took to gain or lose specific folds in the class i and change their respective counts from M_i to P_i. The expectation value over these individual terms is the expected evolutionary time sampled from the overall protein domain compositions. Then the term $\langle\log\frac{P_i}{M_i}\rangle_p - \log\frac{P}{M}$ reflects the *diversification* of the PDC of p. Analogously the same applies to the q-term. As the PDCs are (simplified) descriptors of the organism's proteome we propose that the Jensen-Shannon divergence as used here shows also the diversification of the overall proteomes/protein domain composition.

Therefore $H_{\mathrm{JS}}(p,q)$ can be regarded as the expected divergence time of the *composition of the proteomes* of organisms p and q with respect to an alleged ancestor m. Besides the *a priori*, information theoretically driven motivation of using $H_{\mathrm{JS}}(p,q)$ we can therefore further rationalize about $H_{\mathrm{JS}}(p,q)$ as an indicator of time scales between evolutionary branching events.

Clearly if any two organisms have the same number of superfamily folds i (read $\exists_i\, P_i = Q_i \Longleftrightarrow M_i = P_i \wedge M_i = Q_i$) then this particular superfamily does not contribute to the distance; thus the Jensen-Shannon-entropy incorporates to some extent the idea of previous studies, which stressed the importance of shared folds.

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

4.2 Summary

In this paper we have motivated a new measure of similarity of proteomes that can be rationalized by information theory. We have shown this measure to be a meaningful refinement of previous approaches to whole-genome phylogeny. The advantages in comparison to – say – 16S-RNA-phylogenies are manifold: 1) the natural unit of evolutionary dynamics was argued to be protein domain creation/deletion/'invention', thus making much longer time scales accessible in comparison to a dynamics based on single nucleotide differences; 2) our method also proved to be more robust against variations in the distances that might occur due to errors in the used HMMs among others; 3) besides its motivation from information theory our model can also be justified from a simplified model of evolutionary dynamics; 4) another major advantage is that there is no need for any sequence alignment.

The last property is even more relevant as the quest for 'the correct' alignment procedure is still an open and sometimes troublesome issue in the field of alignment based phylogenies (Morrison and Ellis, 1997; Martin et al., 2007; Rokas, 2008) – it need *not* be addressed when applying our protocol to genomic data. Alignments are only involved in the suggested procedure insofar as they are implicitly contained within the used Hidden Markov Models, but these are not error-prone and remedy the alignment problem therefore.

Conceptually proteomic and other protein-based phylogenies incorporate the effect of various evolutionary "operators" of large complexity (e.g. selection and horizontal gene transfer). Methods focused solely on genomic subsequences, on the other hand, put more emphasis on mutational events, thus they have to map the above mentioned complex mechanisms into single nucleotide changes. Sequence-based approaches based on algorithmic information theory and Kolmogorv complexity are conceptually able to overcome this problem. The complexity measures are, however, not computable (Li and Vitányi, 1997; Kolmogorov, 1965; Solomonoff, 1964a; Solomonoff, 1964b); they must rather be approximated.

Horizontal gene transfer in particular still poses a great challenge to all algorithmic approaches in phylogenetics (Wolf et al., 2002; Snel et al., 2005).

4.3 Outlook

The search for the 'perfect' phylogenetic tree construction algorithm, based on distance matrices, is still ongoing (Woolley et al., 2008). In the future we want to combine our distance measure with established tree construction algorithms besides neighboring-joining and other greedy approaches. The application of global optimization protocols (Hamacher, 2007a; Hamacher, 2006; Wenzel and Hamacher, 1999; Hamacher, 2007b) on e.g. weighted least square minimization (Makarenkov and Leclerc, 1999; Makarenkov, 2001) can in principle provide for a better quality of the mapping distance matrix $\longrightarrow$ phylogenetic trees. The combination of this idea with the distance measure introduced above promises to be even more powerful.

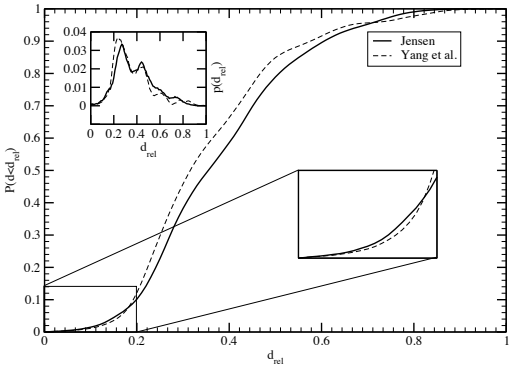

Figure 8: Cumulative distribution $P\left(d_{\mathrm{rel}}^{i,j} < d\right)$ of the relative distances $d_{\mathrm{rel}}^{i,j} := \frac{d^{i,j} - \min\left(d^{i,j}\right)}{\max(d^{i,j}) - \min(d^{i,j})}$ for both methods and all 698 species pairs (i, j). Here d denotes the respective distance measure. The lower inset shows a blow-up of the lower portion of the cumulated distribution, while the upper inset is the distribution itself – one can clearly see the greater richness in structure (larger variation in the slopes) of the distribution of the distance measure by Yang et al., thus indicating more and steeper minima (compare to inset).

ACKNOWLEDGEMENTS

KH gratefully acknowledges financial support by the Fonds der Chemischen Industrie through the program Sachkostenzuschuß für den Hochschullehrernachwuchs. KH gratefully acknowledges support by D. Quandt in the usage of the TreeGraph-package, with which some of the phylogenetic trees in this paper were drawn. The author is grateful for W. Weber's hospitality at the Technical University Dortmund, where this paper was partially written. KH thanks J. Stolze for helpful comments on the paper.

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

Supporting Files Supporting Figure S1 — Comparing Jensen Information Distances for Hidden Markov Models with/without threshold: Scatter Plot of the Jensen Information Distances for all 698 species for the distributions of domains without a threshold and a threshold of $E \leq 10^{-4}$.

Competing interests.
The author declares that no competing interests exist.

REFERENCES

http://bioserver.bio.tu-darmstadt.de/Phylogeny

Burstein, D., Ulitsky, I., Tuller, T., and Chor, B. (2005). Information theoretic approaches to whole genome phylogenies. In *RECOMB*, pages 283–295.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452:745–749.

Ekman, D., Björklund, Å. K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *J. Mol. Biol.*, pages 1337–1348.

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Trans Info Theo*, 49:1858–1860.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166.

Fong, J. H., Geer, L. Y., Panchenko, A. R., and Bryant, S. H. (2007). Modeling the evolution of protein domain architectures using maximum parsomony. *J. Mol. Biol.*, pages 307–315.

Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y., and Nishikawa, K. (2007). A Tree of Life Based on Protein Domain Organizations. *Mol. Biol. Evol.*, 24(5):1181–1189.

Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. *Proteins: Structure, Function, and Genetics*, 33:518–534.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol*, 313(4):903–919.

Grosse, I., Bernaola-Galvan, P., Carpena, P., Romain-Roldan, R., and Oliver, J. e. (2002). Analysis of symbolic sequences using the jensen-shannon divergence. *Phys Rev E*, 65:041905.

Hamacher, K. (2006). Adaptation in stochastic tunneling global optimization of complex potential energy landscapes. *Europhys. Lett.*, 74(6):944–950.

Hamacher, K. (2007a). Adaptive extremal optimization by detrended fluctuation analysis. *J.Comp.Phys.*, 227(2):1500–1509.

Hamacher, K. (2007b). Energy landscape paving as a perfect optimization approach under detrended fluctuation analysis. *Physica A*, 378(2):307–314.

Hamacher, K. (2007c). Information theoretical measures to analyze trajectories in rational molecular design. *J. Comp. Chem.*, 28(16):2576–2580.

Hamacher, K., Hübsch, A., and McCammon, J. A. (2006). A minimal model for stabilization of biomolecules by hydrocarbon cross-linking. *J. Chem. Phys.*, 124(16):164907.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23(2):254–267.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):1–7.

Li, M., Badger, J., Xin, C., Kwong, S., and Kearney, P. e. (2001). An information based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17:149–154.

Li, M., Chen, X., Li, X., Ma, B., and Vitanyi, P. (2004). The similarity metric. *IEEE Trans Info Theo*, 50:3250–3264.

Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151.

Lund, O., Nielsen, M., Lundegaard, C., and Brunak, C. K. S. (2005). *Immunological Bioinformatics*. MIT Press, Cambridge.

MacKay, D. (2004). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2. edition.

Makarenkov, V. (2001). T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668.

Makarenkov, V. and Leclerc, B. (1999). An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *J. Class.*, 16(1):3–26.

Mantaci, S., Restivo, A., and Sciortino, M. (2008). Distance measures for biological sequences: Some recent approaches. *Int J Approx Reasoning*, 47:109–124.

Martin, W., Roettger, M., and Lockhart, P. J. (2007). A reality check for alignments and trees. *Trends in Genetics*, 23:478–480.

Morrison, D. and Ellis, J. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol Biol Evol*, 14(4):428–441.

Otu, H. and Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19:2122–2130.

Citation: K. Hamacher, "Protein Domain Phylogenies", Proceedings of BIOINFORMATICS 2010, p. 114-122, ISBN 978-989-674-019-1, A. Fred, J. Filipe, H. Gamboa (eds.)

Panayiotis V. Benos, Alan S. Lapedes and Gray D. Stormo (2002). Probabilistic Code for DNA Recognition by Proteins of the EGR family. *J. Mol. Biol.*, 323:701–727.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36(1):541–562.

Rokas, A. (2008). GENOMICS: Lining Up to Avoid Bias. *Science*, 319(5862):416–417.

Snel, B., Huynen, M. A., and Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.*, 59(1):191–209.

Solomonoff, R. J. (1964a). A formal theory of inductive inference. *Information and Control*, 7:1–22.

Solomonoff, R. J. (1964b). A formal theory of inductive inference. *Information and Control*, 7:224–254.

Soltis, P. S. and Soltis, D. E. (2003). Applying the bootstrap in phylogeny reconstruction. *Statist. Sci.*, 18(2):256–267.

Tekaia, F., Lazcano, A., and Dujon, B. (1999). The Genomic Tree as Revealed from Whole Proteome Comparisons. *Genome Res.*, 9(6):550–557.

Wenzel, W. and Hamacher, K. (1999). A Stochastic tunneling approach for global minimization. *Phys. Rev. Lett.*, 82(15):3003–3007.

Wilson, D., Madera, M., Vogel, C., Chothia, C., and Gough, J. (2007). The superfamily database in 2007: families and functions. *Nucleic Acids Res*, 35(Database issue):308–313.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc. Nat. Acad. Sci.*, 97(15):8392–8396.

Woese, C. R. (2002). On the evolution of cells. *Proc. Nat. Acad. Sci.*, 99(13):8742–8747.

Wolf, Y., Rogozin, I., Grishin, N., and Koonin, E. (2002). Genome trees and the tree of life. *Trends in Genetics*, 18(9):472–479.

Woolley, S. M., Posada, D., and Crandall, K. A. (2008). A comparison of phylogenetic network methods using computer simulation. *PLoS ONE*, 3(4):e1913.

Yang, S., Doolittle, R. F., and Bourne, P. E. (2005). Phylogeny determined by protein domain content. *Proc. Nat. Acad. Sci.*, 102(2):373–378.

Zhang, Y. and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proc. Nat. Acad. Sci.*, 102(4):1029–1034.