A pipeline for finding characteristic amino acid substitutions for cold-adapted proteins

Gisle Sælensminde [*†], Nils-Peder Willassen [‡] and Inge Jonassen [,§]

April 18, 2005

1 Abstract

Enzymes adapted to a cold environment often have very similar structure to mesophilic homologs, and most significant changes in catalytic activity at low temperature is due to more subtle substitutions. We have made a computational pipeline based on the Pfam protein domain families, where we build subset families and estimating the amino acid substitutions that have happened when adaptations to a cold environment have taken place. We also take into account where in the structure the amino acid substitutions happen. Especially, we think that solvent accessible vs the buried residues should be distinguished, since they usually have very different roles in the protein stability and function. We hope that this will tell us something about which methods that are most dominant for cold adaptations. The Pfam subset consists of about 310.000 sequences in 3859 families. Of these 1536 has at least one pdb-structure resolved for the domain.

2 Introduction

Enzymes from psychrophilic (cold adapted) organisms often keep a high catalytic activity at low temperatures, while the activity of homologs from mesophilic species (adapted to moderate temperatures) drops to very low values. This seems to happen without any major structural changes to the protein fold. Studies of protein families with sequences from both mesophilic and psychrophilic species suggest several different patterns for adaptation. Since many enzymes have domain movements during the catalytic activity, the protein have to be more flexible to compensate for slower molecular movements. Another mechanism observed is more favorable charges on the surface near the active site, so the ligands can bind to it more easily. In this project we study protein domains from the Pfam [1] database, and build phylogenetic tree of the sequences where we know the growth temperature for the species. Pfam is not really a sequence family database, but represents protein domains. We selected Pfam, because it have quite good alignments, and unlike whole-protein databases like hobacgen, Pfam do not have sequence parts that is unrelated to the other sequences in the family.

3 Methods

The pipeline have five steps (Figure1) The first step is to extract the sequences from Pfam where we know the growth temperature of the speecies. In the next step we build a phylogenetic tree for the sequences remaining in the family. Furthermore we reconstruct the ancestral sequence and identifies the branches that represents cold adaptations. Based on this, we can find the substitution patterns for sequences adapting to a cold environment. Quite parallel to this we extract features from the protein structures, so we can use this to distinguish between different parts of the structures that may have very different substitution patterns.

[*] author for correspondance, gisle@cbu.uib.no

[†] Computational Biology unit(CBU), BCCS, University of Bergen, Norway

[‡] Institute of medical biology, Faculty of Medicine, University of Tromsø, Norway

[§] Computational Biology unit(CBU), BCCS, University of Bergen, Norway

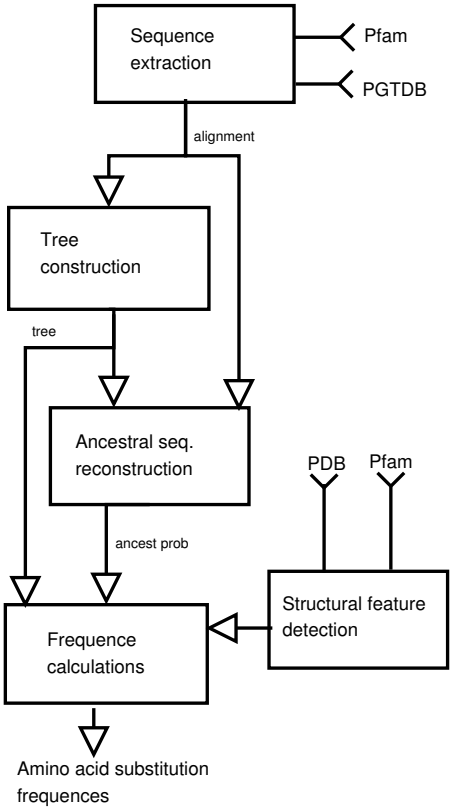

Figure 1: The pipeline for computing the amino acid substitution rates for proteins adapting to a cold environment. First sequences with known growth temperature is extracted. Then the tree and ancestral sequence probabilities are computed. This is held together with structural information, like surface exposure to find the substitution rate in different structural areas.

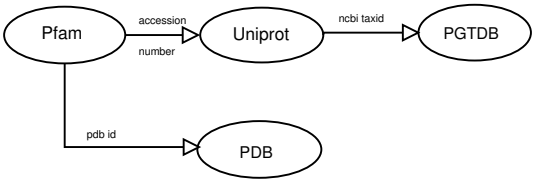

Figure 2: The cross references used in the pipeline. The temperature for each sequence is found by looking up the NCBI taxonomy id in uniprot, to find the species, and then look up the temperature by taxonomy id in PGTDB. Protein structures are found by using the PDB cross references in PFAM. Secondary structures and active site location are stored directly in Pfam.

3.1 Extraction of sequences

Temperature environment for the different species are not easily available for all species. For bacteria and archea, the procaryotic growth temperature database (PGTDB)[5] contains temperature information for a greater number of species. How many depends on how one counts, since the information is not complete for all species. For the purpose of this study, about 1000 species are available. This database contains range (upper, lower) for optimal growth temperature, as well as common lab temperature condition for the organism. The species are identified by the NCBI taxonomy id[2]. Since we are interested in gaining knowledge about temperature adaption, we are only interested in those species for which we know the optimal temperature environment, so we extracted the subset of Pfam that represent species in the PGTDB. Since Pfam has no cross reference to taxonomy ids, we needed to go the indirect route, and reference the swissprot/trembl [9] entry for the sequence, that contains the taxonomy ids, that could be matched against PGTDB (Figure 2). The pfam domains that has at least three domain instances left after the extraction were kept.

3.2 Phylogenetic trees

We are basing the phylogenetic tree reconstruction on the protein sequences, rather than DNA sequences. Methods like Maximum Likelihood(ML) and Bayesian inference in most cases gives better trees than simple distance based methods like neighbor-joining. On the other hand espe-

cially ML-methods are quite slow, so we have chosen to nevertheless use neighbor-joining rather than a more sophisticated method in order to spend less time on computations.

3.3 Ancestral sequence reconstruction

For ancestral sequence reconstruction we use a parsimony method [3]. While ML or bayesian methods may give more accurate results, the parismony method is fast, which is important when we are analyzing many protein families. Since the method requires rooted trees we finds the root by using the midpoint algorithm.

Most ancestral sequence reconstruction methods try to reconstruct the most likely amino acid in each position. This will more often be a frequently occuring amino acid than an infrequent one, and thus the most frequent amino acids will be even more frequent in the reconstructed sequences, especially near the root. Some of this bias can be avoided by replacing the amino acids in the ancestral sequences by probabilies of the different amino acids, given the present sequences, but there is apparently still bias toward frequent amino acids with the parsimony method, even when using prababilities, so we need to take that into account when analyzing the results.

When using amino acid probabilies like this, we cannot tell for certain that a particular amino acid mutated to another along a branch, but get probabilities for the different mutations, and we are summing these probabilies instead of counting mutations.

3.4 Identifying cold adaptation branches

Since the temperature environment is generally not known for the ancestral species, or at least not readily available, we have to infer the temperature environment based on the temperature of the present day species, which is associated with the leaves in the tree. There is no obvious way of doing this, but we have chosen to use a parsimony approach (Figure 3). We group the species in three categories; psychrophilic, mesophilic and thermophilic, using a character based approach, meaning that we reconstructing the temperature conditions for the species as they should have been amino acids or bases. (e.g psychrophilic instead of say alanine). In fact we are using the exact same method for temperature reconstruction as for reconstruction of one site (column) in the multiple alignment.

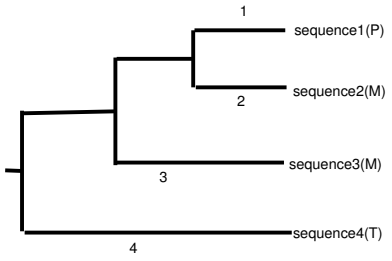

Figure 3: Branch temperature assignment: The parsimony principle that the model that gives the least changes are assumed. In the tree above, sequence 1 is psychrophilic, sequence 2 and 3 is mesophilic, while sequence 4 is thermophilic. The least number of temperature adaptations will happen when branch 1 and 4 change

The problem with inferring temperature environment this way, is that the sequence databases have a bias towards mammalian sequences and pathogenic bacteria. The former is filtered out, but the latter may cause some branches to be falsely identified as a cold adaptations. We will analyze the amino acid transition probabbilities where we have high transition probability for one transition (mesophilic to psychrophilic, mesophilic to mesophilic etc). Currently we are using branches with probability for cold adaptation of 0.95 or more.

3.5 Finding structural elements

For 1536 of the 3859 families there is at least one PDB-structure available. This makes it possible to identify structural features of the domain/family. The most relevant features are solvent exposed surface vs core. Other features, like secondary structures and active site location, and interface between different protein chains [1] can also

[1] This can be difficult to get precise, since the knowledge about whether the protein is multi-chained is often disputed. Even though the crystal structure forms a multi chain complex, this may not be the case in vivo.

be interesting. In order to find the solvent exposed surface area, we are using the msms program [10]. This is a simple geometry-based program that calculates the surface area based on a list of atomic coordinates and radii. We are mainly interested in the exposure of the side chains, since that is what differs between the amino acids. The interface area can be calculated by calculating the surface area with all chains, and the chains independently, and then look at the differences. The areas that are more exposed when the surface exposure of the chains are calculated independently for each chain, are assumed to be part of an interface area. Currently we have not finished the implementation of surface exposure detection or other structure-based metods, so this is not presented in the results.

3.6 Output

The output of the method is for each triplet [from temperature class, to temperature class, structural class], a $20*20$ substitution matrix, where each element $M_{i,j}$ is the number of substitutions from amino acid i to amino acid j.

Residue	P->M	M->M	Occ
A	0.84	0.89	2.08
R	0.81	0.95	1.08
N	1.50	1.18	0.69
D	0.89	0.95	1.08
C	2.30	1.50	0.19
Q	1.34	1.23	0.69
E	0.67	0.85	1.17
G	0.81	0.92	1.71
H	1.58	1.25	0.44
I	1.18	1.08	1.27
L	0.62	0.80	2.07
K	1.07	0.99	0.86
M	1.98	1.41	0.48
F	1.31	1.12	0.78
P	0.60	0.84	0.87
S	1.42	1.18	1.11
T	1.25	1.13	1.06
W	1.44	1.19	0.22
Y	1.16	1.07	0.56
V	1.03	1.01	1.54

Figure 4:

4 Results

We tested the pipeline on 191 families from Pfam . The result of this is presented in figure 4. In columns 2 and 3, we show for each amino acid the ratio of residues that mutates to this amino acid, divided by the number of residues that mutates from it to some other amino acid. Column 2 of the table shows the mutations along branches leading from a mesophilic to a psychrophilic node, while column three is for mutations beteen mesophilic nodes. Since column 3 for mutations between sequences from two mesophilic species, the values should have been about 1.0 with an unbiased method. It quite obvious that this is not the case, instead the frequences seems to be correlated with the amino acid frequencies. For comparision the amino acid frequencies divided by 1/20, the expected frequency if all amino acids had the same frequency. In fact there is a negative correlation of -0.70 between the amino acid frequencies and the mesophilic mutation rate, and this indicates a very strong bias toward the amino acid frequency. The effect is probably due to the by a bias in the parimony method for ancestral reconstruction that we use. This method may have a tendency to over- represent the amino acid with high frequency in the present day species in the ancestral sequences. If you follow the branches from the common ancestor toward the leaves of the tree, you will then see that frequent amino acids “disappears”.

The problem with this is that it makes it hard to compare that values from mesophilic and psychrophilic species, since an increase or decrease in frequency compared to mesophilic will cancel out with the negative correlation to amino acid frequency, and since the relationship between amino acid frequency and mutation frequence is possibly non-linear, it is very hard to interpret these values.

In fact the output of this analysis is quite in conflict with the literature in the field [7] [4]. Cold-adapted proteins is belived to have less aromatic and more aliphatic and charged amino acids. This is quite the opposite of what we see, but is most likely due to the amino acid frequences.

5 Discussion

It is clear that the ancestral reconstruction is quite biased, and of the next steps must be to add a more sophisticated ancestral reconstruction method that avoids biases. The fastml program [8], is a good candidate for a better ancestral reconstruction program, but it requires quite a bit more cpu-time than a parsimony reconstruction, and this can be a bottleneck when we are calculating trees for thousand of families, but since we have severe problems with the method as we use it now, we will look into this. In addition we should evaluate this or alternative methods by simulation, so we can detect methodical biases as we apparently have in this case.

We have not evaluated the quality of the phylogenetic trees we are using, but we are using a fast and inaccurate method, so we may change to MrBayes [6], that is a Bayesian method. We expect ML tree reconstruction methods to be too slow for our purpose.

Of structural analyses, the surface exposure algorithm has higest priority. Other structural features can be analysed, like forming and dissappearance of hydrogen bonds, salt bridges etc. These structural features may not as easily fall into the framework as surface exposure, but since it is belived that the number and position of these play an important role for adaptation both to cold and hot environments. It is also belived that there are more important changes around the active site, so analyzing the residues in a radius from the active site will probably be valuable as well.

References

[1] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna2, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeat s, and Sean R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32:D138–D141, 2004.

[2] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. Genbank: update. *Nucleic Acids Research*, 32:D138–D141, 2004.

[3] Walter M Fitch. Toward defining the course of evolustion: Minimum change for a specific tree topology. *Systematic zoology*, 20:405–416, 1971.

[4] D. Georlette, V. Blaise, T. Collins, S. D'Amico, E. Gratia, A. Hoyoux, J.C. Marx, G. Sonan, G. Feller, and C. Gerday. Some like it cold: biocatalysis at low temperatures. *FEMS Microbiology Reviews*, 29:25–42, 2004.

[5] Shir-Ly Huang, Li-Cheng Wu, Han-Kuen Liang, Kuan-Ting Pan, Jorng-Tzong Horng, and Ming-Tat Ko. Pgtdb: a database providing growth temperatures of prokaryotes. *Nucleic Acids Research*, 20:276–278, 2004.

[6] Huelsenbeck JP and Ronquist F. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–755, 2001.

[7] Sandeep Kumar and Ruth Nussinov. Different role of electrostatics in heat and in cold: Adaption by citrate synthase. *ChemBioChem*, 5:280–290, 2004.

[8] T. Pupko, I. Pe'er, D. Graur, M Hasegawa, and Friedman N. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics*, 18:1116–1123, 2002.

[9] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, and Yeh LS. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32:D115–D119, 2004.

[10] M. F. Sanner, Olson A. J., and J. C. Spehner. Fast and robust computation of molecular surfaces. *ACM 11. Symposion on compuattional geometry, preceedings*, pages C6–C7, 1995.