DeltaProt: Molecular comparison of proteins based on sequence alignments.

A Matlab© companion Toolbox (v.2.1):
http://www.math.uit.no/bi/deltaprot/

This software can be used freely for academic, non-commercial use.
Note: Matlab Statistical Toolbox is used in some functions.

Steinar Thorvaldsen[1], Tor Flå[1] and Nils P. Willassen[2]
[1]University of Tromsø, Faculty of Science and
[2]Norwegian Structural Biology Centre
9037 Tromsø - Norway.

steinar.thorvaldsen@uit.no

Abstract. We present implementations of comparative analyses and statistical trend-tests that are useful when the protein sequences in multiple alignments can be divided into two or more subgroups based on a known phenotype assignment. The development of *DeltaProt* has been motivated by the need to create a flexible software platform that enables statistical analyses of orthologous proteins with special environmental preferences (temperature, pH, salt concentration or pressure). We also provide procedures to plot the output from these tests for visualisations. The algorithms have been successfully applied in research on extremophile proteins.

1. Installation

To run the program *DeltaProt* you should make the following steps:

1. Download the program to your local PC and install (unzip) it in a fresh folder like …\deltaprot
2. Run Matlab v. 7.1 or later
3. Set working directory to ...\deltaprot (root directory of the toolbox)
4. Put your alignment(s) data in a subdirectory like ...\deltaprot\data
5. Modify one of the main program scripts to provide access to your data.

To run the programs, it is necessary that a subdirectory for data exists in the working directory. This directory must contain the sequence alignment(s) in fasta file format.

2. Features

Comparative bioinformatics is an emerging field in which knowledge coming from sequence alignments are analyzed to find out more about the properties of living systems. The novelty of analytical methods used and the special nature of the data used, require the development of new software tools. DeltaProt is a Matlab™ (versions 7.1 or later) toolbox developed for the comparative analysis of alignments. It was designed as a Matlab companion with a set of functions and algorithms for modeling multiple alignments of proteins by a range of statistical methods. The approach has been successfully applied in research on extremophile organisms (Thorvaldsen *et al*., 2007; Thorvaldsen and Ytterstad, 2009), and has special relevance for membrane proteins, since it is very difficult to obtain structural models of this important group. This is done by using comparative statistical methods on extremophile proteins versus ortholog genes from organisms of normal habitats. The approach is also suitable for high-throughput sequence analysis.

The protein sequences must be aligned by using one of the available alignment programs prior to import into DeltaProt. DeltaProt may read alignments in standard FASTA file format. Each alignment may also contain information on secondary structure of the amino acids, and the accessible surface area (ASA), either obtained from a template structure, or from a sequence based prediction program (Adamczak et al. 2004). If the secondary and/or 3D structure is known, or may be predicted, the analyses can be performed in each of these regions.

The toolbox consists of a set of statistical routines with a variety of modeling functions. We consider both the amino acid sequence *compositions*, and the *substitution* patterns, to determine whether there are underlying trends that explain the observed variation between the phenotypic groups to be analyzed. More than 80 different *physicochemical properties* of the amino acid may also be applied in order to reduce the sequence alphabet to measurements. Each situation is analyzed by appropriate statistical methods. Table 1 shows an overview of the statistical models and tests available in DeltaProt.

The statistical models have a very flexible design. In some cases data from several different taxonomic orders may be available, and this information can be utilised, by a *stratified* approach, where the biological order is included in the model. The stratification of the subjects into disjoint sets increases the power of the test to detect association, because like subjects are compared to like subjects.

The approach assumes statistical independence among the sequences samples in each group at each strata, but it may be modified to treat phylogeneticly dependent sequences within the groups. When there are multiple dependent samples in a group, we then compute the mean values in each group at each stratum, and apply the statistical test to the means as representative observations.

DeltaProt outputs summary statistics. These summary statistics include mean, standard deviation and p-values from the comparative tests. When multiple hypothesis

tests are carried out, significance levels should be adjusted to account for the increased probability of false positives. For multiple test correction, the False discovery rate analysis (FDR) of Benjamini and Yekutieli (2001) may be applied.

The differences between the phenotypic groups may also be compared graphically. Amino acid composition, substitution patterns and physiochemical properties along the sequences may all be visualized.

3. Overview

The program suite consists of two main program scripts and about 30 procedures. The main scripts should be modified to provide access to new alignment data.

Table 1

File name	Brief description
DeltaProt1.m [1] DeltaProt.m [N]	Main program scripts
readfasta.m aa2int.m dssp2int.m wasa2int.m	Reads text files (alignments) in fasta format Transforms symbols of amino acids Convert symbols for secondary structure to numbers Convert 3D predictions to numbers expressing three levels
findSeqSimilarity.m findSeqSiteVar.m	Finds overall similarity between sequences in alignments Finds variations (Var≥2) and conserved sites(Var=1)
aaCount.m aaAnova1Pvalues.m aaAnova2Pvalues.m [1] aaFreq.m aaFreqPlot.m aaDeltaFreq.m aaDeltaFreqPlot.m	Calculates counts of amino acids Runs one-way unbalanced ANOVA test Runs two-way unbalanced ANOVA test Calculates frequencies and extracted previously from file, and Plots resulting frequencies Calculates compositional changes between sequences Plots compositional changes
substPairsCount.m substPairsPvalues.m substPairs2bin.m FisherExtest.m* chi2Tests.m* MantelHaenTest.m*, [1] substPairsPlot.m substCountSort.m substPvalueSort	Calculates substitutions between aligned sequence pairs Runs the appropriate statistical test Reduce the full substitution matrix to fewer categories Fisher's exact test with mid-P-values Chi-square tests (Read-Cressie, Pearson or Log Likelihood) Stratified Mantel-Haenszel test Plot the substitution matrix results Sort the substitutions by numbers Sort the substitutions by P-values
filterAlignedseq.m propAnovaPvalues.m propNormalityTest.m* propPairedPvalues.m	Reduces sequence data to property data. Runs one-way ANOVA test. Optional Box-plots. Data-adaptive goodness of fit to normality Runs paired trend-tests (t-test or Wilcoxon). Optional plots.

propRegress.m [1] propRegressKendall.m [N] cumKendallTest.m* FDR.m	To detect trends in ONE protein by parametric linear regression To detect trends by non-parametric linear regression. Implemented cumulative Mann-Kendall trend test False Discovery Rate for multiple test correction
aa_prop60.xls	Excel-file with the physicochemical properties

* Implemented statistical test in *DeltaProt*.
[1] Appropriate for *one* protein family when sequence data from several taxonomic orders may be utilized as strata in the statistical analysis. [N] Appropriate for a set of *several* alignments.

DeltaProt was developed with Matlab and is distributed in the standard Matlab language, making it compatible with multiple platforms (Windows, Mac, Linux and Unix). It requires at least Matlab™ R14 (2005). Some functions require the Matlab™ Statistics Toolbox. DeltaProt was developed as an user-extensible environment that facilitates the flexibility and exchange of analytical methods in the types of problems that arise in comparative genomic studies. It provides access to implementation details and encourages modification and extension of capabilities.

DeltaProt may be used to analyze *one* protein family, or it may be used to analyze alignments from *several* protein families from a set of genomes, to look for common adaptive trends at the molecular level. In the first case ordinary linear regression is applied, and the second case may be performed by non-parametric regression based on a cumulative Mann-Kendall trend test (Thorvaldsen and Ytterstad, 2009).

3. A sample session with *one* protein family

This script calculates statistical P-values for monotonic trend for three ordered groups of sequences in one alignment:

```
% DeltaProt1.m   A comparative analysis of ONE aligned protein sequence.

%---------------------- Read and converts data -----------
aa_seq = dir('./data1/*.fas'); % reads the filname from data directory
% The three temperature habitats (phenotypic groups):
t31_40=[1 2 3 4 5 6 17 18 19]; %mesophiles
t21_30=[7 8 9 10 11 12 20 21 22 23 24 25 34]; %intermediate
t1_20=[13 14 15 16 26 27 28 29 30 31 32 33 35]; %psycrophile
top_lines=4; %line 1..4 in alignment reserved for 2D/3D information
% L is the number of lines in the alignment file, n is the length,
% S is the character sequence array, and names are the headers:
[L, n, S, names] = readfasta(strcat('./data1/',aa_seq(fileNo).name));
Sint=aa2int(S(top_lines+1:L,:)); % integer number representation
% Find the variation and conserved sites (Var=1) along the sequences
siteVar = findSeqSiteVar(Sint);
% Computational filters added at the top or bottom of the alignment:
level2D=dssp2int(S(1,:)); % secondary structure(Alpha; Beta Loop)
level3D=wasa2int(S(3,:)); % internal/external amino acids(0..9)
% We use cell arrays as a storage mechanism for dissimilar size of data:
Sgrouped= {Sint(t31_40,:); Sint(t21_30,:); Sint(t1_20,:)};
```

```
%---------------------- I. Compute and plot aa-composition---------------
% Structural 3D location:
[aaComp3D,aaStd3D,aaTotal3D] = aaFreq(Sgrouped, level3D);
aaFreqPlot(aaComp3D, aaStd3D, 3);

%----------------- II. Compute and plot aa-substitution matrix---------
substMP3D = substPairsCount(Sint(t31_40,:), Sint(t1_20,:), level3D,1,1);
substMM3D = substPairsCount(Sint(t31_40,:), Sint(t31_40,:, level3D,1,0);
PvalueMPvsMM3D = substPairsPvalues(substMP3D, substMM3D);
substPairsPlot(substMP3D,PvalueMPvsMM3D, 0.05);

%---------- III. Compute group trends based on properties--------------
% To apply the regression procedure we first have to modify the data
% format for temperature group (1=mesophiles, 2=interm, 3=psycrophile):
X(:,1)=[1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3
3 2 3]'; %temp habitat 1-3
% We look for significant physiochemical trend changes in the groups:
[b,prop_P,prop_name]=propRegress(Sint, X, level3D, siteVar);
```

Each function is documented in the upper part of its script.
On a Pentium IV 1.6 GHz with 512 MB RAM running Windows XP, this script uses less than one minute with 30 sequences in the alignment and 60 properties.

Some of the output graphics from the script are shown below.

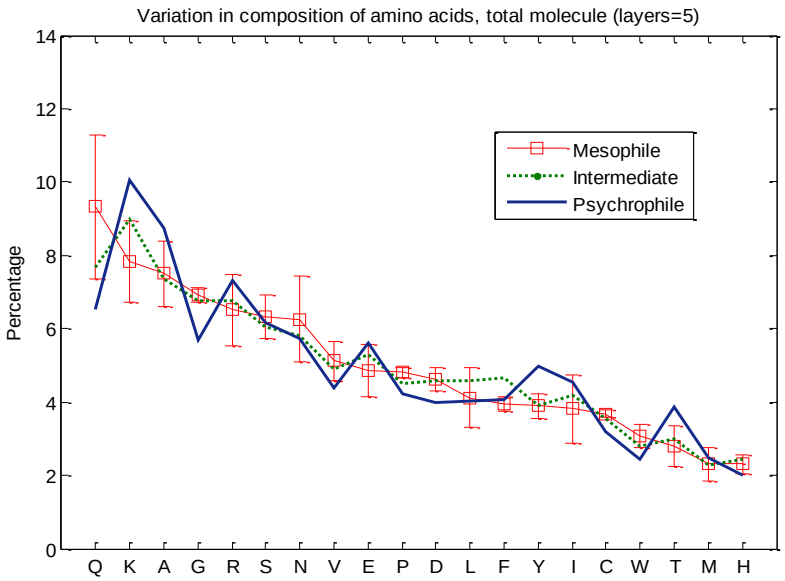

Figure 1: Plot of amino acid composition observed in the three phenotypic groups. Error bars represent the empirical standard deviations of the mesophile group. The Anova analyses show that the decrease of amino acids Q and G, and the increase of K and Y, can be significantly related to temperature.

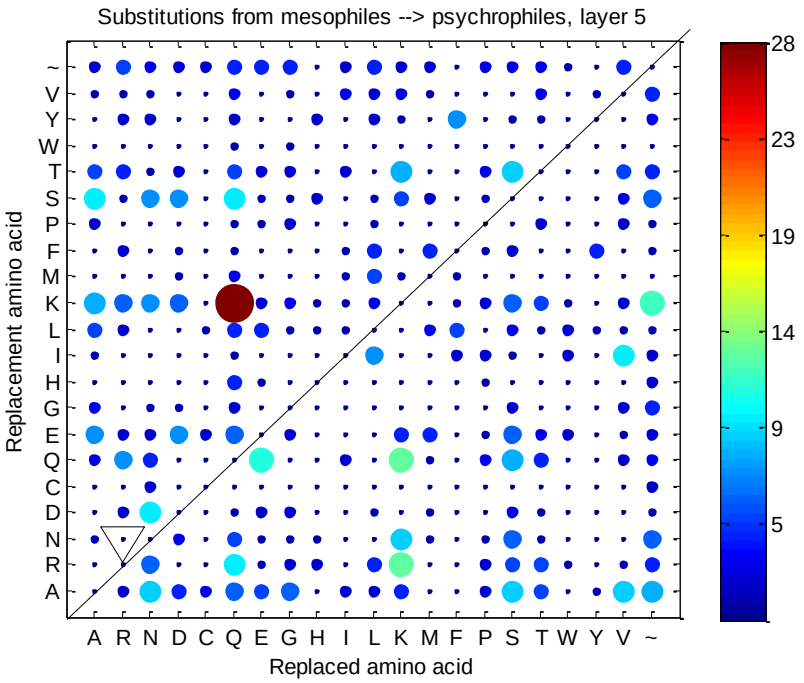

Figure 2. Visualization of the number of pairwise substitutions observed between two groups. The size and colour of each marker indicates the magnitude of the substitution (see colour-bar). A tilde (~) indicates deletion/insertion. Non-favoured substitutions with p-values < 0.05 in the Replaced (mesophile)→Replacement (psychrophile) direction are marked with a dovnward-pointing triangle.

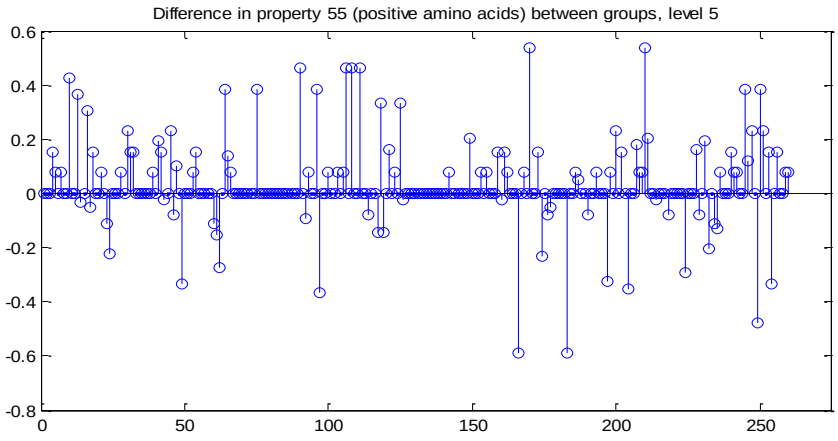

Figure 3. Mean difference in the number of positive amino acids along the sequence. P-value = 0.0003 by the liner regression analysis.

An example session of a stratified analysis of the same dataset, where the biological order is included in the model, is also included in the script *DeltaProt1strata.m* .

4. A sample session with *several* protein families

This script calculates statistical P-values for monotonic trend for three ordered groups of sequences in 25 alignments:

```
% DeltaProt.m: multiple protein comparison, main program sample script.

%---------------------- Read and convert data --------------------------
aa_seq = dir('./data/*.stn'); % reads the filname from data directory
maxNo=size(aa_seq,1); % the number of files in this particular case
M=12; %the number of data lines represented in the fasta-files

% Defining sequence group of all 7 sequences (grouped by temperature):
maxPop=3;
t31_40=[2 3 4]; % the mesophile sequences, nr 1 is a copy
t21_30=[5];
t1_20=[6 7]; % the psychrophile sequences.
extra_lines=5; %line 8..12 reserved for 2D/3D information

%Here dim 2 and 4 must match with data! May be changed:
aaDeltafreq2D = zeros (maxNo,maxPop-1,20,maxlevel+1); % for storing
changes in frequencies
aaDeltafreq3D = zeros (maxNo,maxPop-1,20,maxlevel+1); % for storing
changes in frequencies
proteinBlock = zeros(maxNo,1); % for storing the length of each protein
sequence block
Stotal=[];
siteVartotal=[];
level2Dtotal=[];
level3Dtotal=[];

for fileNo=1:maxNo
    % L is the number of lines in the alignment file, n is the length,
    % S is the character sequence array, and names are the headers:
    [L, n, S, names] = readfasta(strcat('./data/',aa_seq(fileNo).name));
    proteinBlock(fileNo) = n; % length of this alignment
    Sint=aa2int(S(1:L-extra_lines,:)); % integer number representation
    siteVar = findSeqSiteVar(Sint); % finding the site variation
    Sgrouped= {Sint(t31_40,:); Sint(t21_30,:); Sint(t1_20,:)};
    % Computational filters added at the top or bottom of the alignment:
    level2D=dssp2int(S(9,:)); % secondary structure(1,2,3)
    level3D=wasa2int(S(11,:)); % internal/intermediate/external(1,2,3)
    Stotal=[Stotal,Sint]; % store data by extending the arrays in dim 2
    siteVartotal=[siteVartotal,siteVar];
    level2Dtotal=[level2Dtotal,level2D];
    level3Dtotal=[level3Dtotal,level3D];
    aaDeltafreq2D(fileNo,:,:,:) = aaDeltaFreq(Sgrouped, level2D);
    aaDeltafreq3D(fileNo,:,:,:) = aaDeltaFreq(Sgrouped, level3D);

end
%-------------- I. Plot the frequence variation of amino acids---------
% Secondary structure:
%aaDeltaFreqPlot(aaDeltafreq2D);
```

```
% Structural 3D location:
aaDeltaFreqPlot(aaDeltafreq3D);


%---------------------- II. Compute and plot aa-substitution matrix-----
--
% Secondary structure of protein (exsample)
substMP2D = substPairsCount(Stotal(3,:), Stotal(6,:), level2Dtotal, 1);
PvalueMPvsPM2D = substPairsPvalues(substMP2D);
%substPairsPlot(substMP2D,PvalueMPvsPM2D, 0.05);
substPvalueSort(substMP2D,PvalueMPvsPM2D);
substCountSort(substMP2D,PvalueMPvsPM2D);

% 3D structure of protein (exsample)
substMP3D =
  substPairsCount(Stotal(t31_40,:),Stotal(t1_20,:),level3Dtotal, 0, 1);
PvalueMPvsPM3D = substPairsPvalues(substMP3D);
substPairsPlot(substMP3D,PvalueMPvsPM3D, 0.01);
substPvalueSort(substMP3D,PvalueMPvsPM3D);
substCountSort(substMP3D,PvalueMPvsPM3D);
substMM3D =
  substPairsCount(Stotal(t31_40,:),Stotal(t31_40,:),level3Dtotal, 0, 0);
PvalueMPvsMM3D = substPairsPvalues(substMP3D,substMM3D);


%--IIIc. Compute and test poulation properties nonparametric regression-
Sgrouped= {Stotal(t31_40,:); Stotal(t21_30,:); Stotal(t1_20,:)};
% We look for significant physiochemical trend changes:
%[b2D,cumKendall_p2D,prop_name]=
  propRegressKendall(Sgrouped, level2Dtotal,siteVartotal, proteinBlock);
[b3D,cumKendall_p3D,prop_name]=
  propRegressKendall(Sgrouped, level3Dtotal,siteVartotal, proteinBlock);
% For ploting slopes of one property (No 5):
Property5_Surfacelevel3_Allprot_Slope=sort(squeeze(b3D(5,3,:,2)));
figure;
bar (Property5_Surfacelevel3_Allprot_Slope);
```

On a Pentium IV 1.6 GHz with 512 MB RAM running Windows XP, this script uses less than one minute.

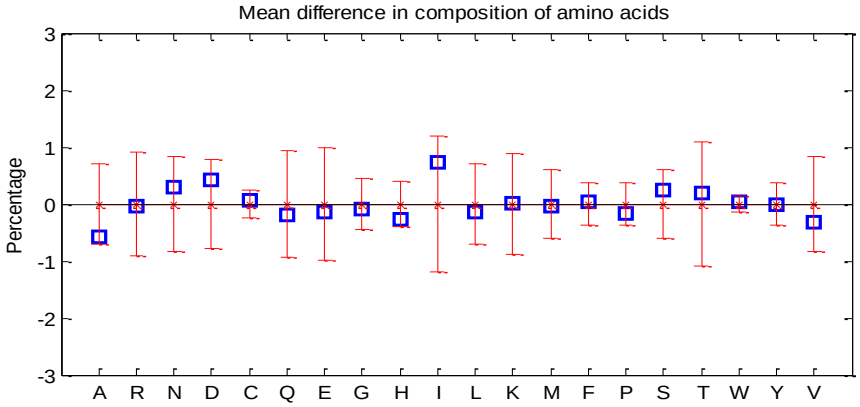

Figure 4. Comparison of the mean amino acid compositional changes calculated on the basis of 25 orthologous proteins observed in the direction from the mesophile to the psychrophile group. Error bars represent the empirical standard deviations. By the cumulative Mann-Kendall trend test the amino acids A, D, and I have a significant change (p-values < 0.05).

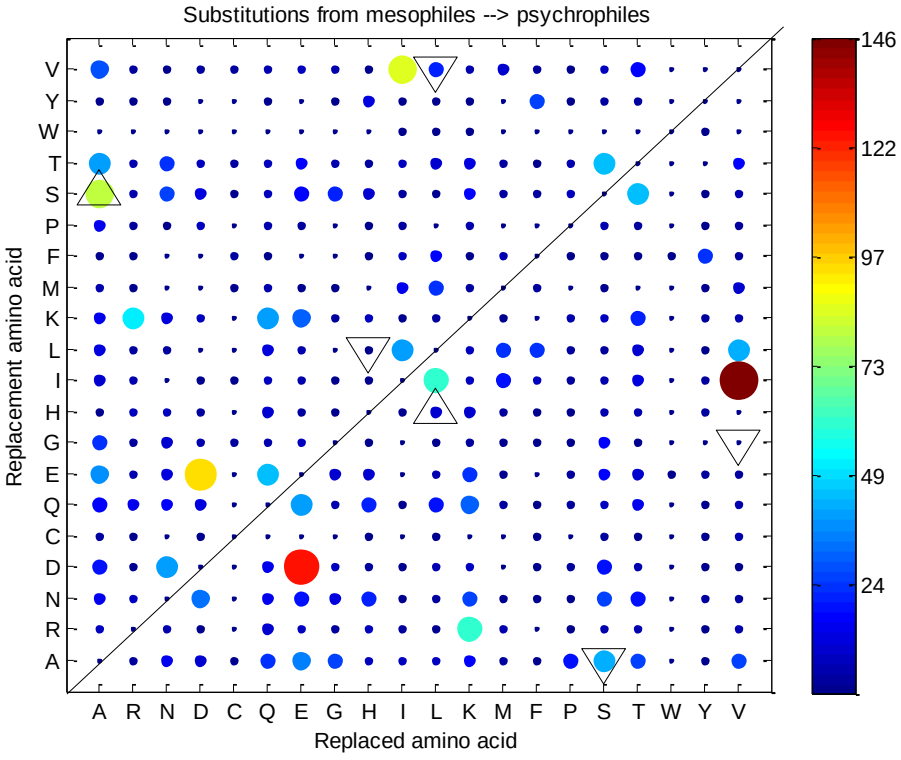

Figure 5. Visualization of the number of pairwise substitutions observed in a comparison of 25 ortholog proteins between two groups. The size and color of each marker indicates the magnitude of the substitution (see color-bar). Favored substitutions (P-value < 0.01) in the Replaced→Replacement direction are marked with upward-pointing triangles, and the non-favored substitutions (P-value < 0.05) are marked with downward-pointing triangles.

DeltaProt generates output statistics like this:

The most biased substitution pairs (SP) at

SP	Forward	Reverse	P_values (with +/- as direction)
HL:	1	11	P= -0.0002090
LV:	22	43	P= -0.0027969
LH:	11	1	P= 0.0034270
SA:	42	82	P= -0.0046187
AS:	82	42	P= 0.0064186
VG:	0	5	P= -0.0086521
VI:	146	85	P= 0.0106786
LD:	6	0	P= 0.0123351
YH:	3	12	P= -0.0179396
DL:	0	6	P= -0.0224965
AH:	2	8	P= -0.0234762
KI:	8	1	P= 0.0268535
EI:	5	0	P= 0.0294613
KT:	12	22	P= -0.0313289
HM:	0	3	P= -0.0339411
WF:	2	0	P= 0.0344180
TY:	4	0	P= 0.0348272
GV:	5	0	P= 0.0353297
QH:	10	21	P= -0.0399307
YT:	0	4	P= -0.0427016

Most frequent substitution pairs (SP) in numbers:

SP	Forward	Reverse	P_value (with +/- direction)
VI:	146	85	P= 0.0106786
ED:	124	95	P= 0.2069409
DE:	95	124	P= -0.7196988
IV:	85	146	P= -0.8114528
AS:	82	42	P= 0.0064186
KR:	61	51	P= 0.7267839
LI:	60	41	P= 0.0437624
RK:	51	61	P= -0.3813850
TS:	45	44	P= 0.7105821
QE:	44	39	P= 0.5418795

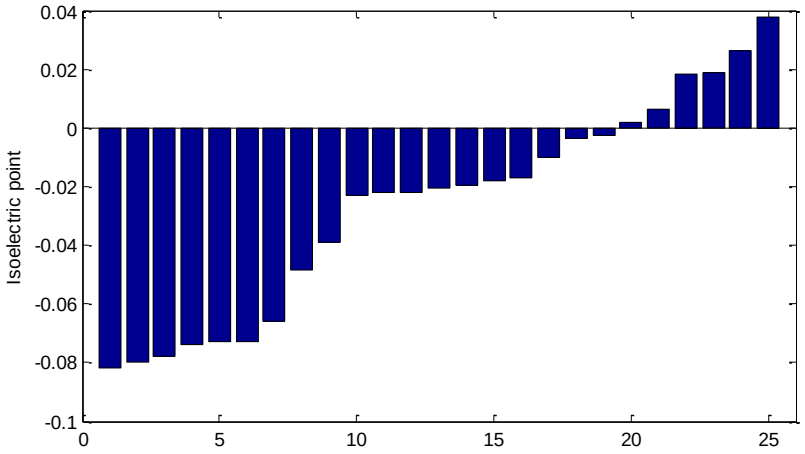

Figure 6. The ranked distribution of slope coefficients for change of isoelectric point at the predicted surface of 25 proteins as estimated from the regression model. Each of the alignments consists of 6 sequences from 3 phenotypic groups. Computed p-value = 0.0067 from the cumulative Mann-Kendall trend test.

An example session of an analysis of 65 orthologous membrane proteins is also included in the toolbox script *DeltaProtMembrane.m* .

5. FAQ

What knowledge of programming is needed?

The users must have working knowledge of programming in Matlab. The code for running DeltaProt is entered through a command line interface, and the code must be modified to access new data.

How big a computer is needed?

DeltaProt runs well on a laptop PC with Matlab installed. The computational requirements depend heavily on the size of the data set. Generally, the toolbox requires high numerical processing capabilities only due to a large data set, i.e. when a whole proteome consisting of more than 1000 alignments, each with 10 sequences, are to be processed.

How should the sequence alignment be made?

Use one of the many alignment programs like BioEdit og T-Coffee. Store the data in standard FASTA file format.

How should the additional structural data be added to the input file?

If the structure is unknown, it may be predicted using a structure prediction program like SABLE (Adamczak et al. 2004). DeltaProt does not provide any Matlab functions that can help users to create a valid input file containing extra information from the

structure predictor. This has to be done manually or by modifying the Python script we have included with the toolbox.

How many sequences are needed?
If you are analyzing only *one* alignment, you probably will need at least 20-25 sequences that are divided in two or more phenotypic groups. If the diversity signal is strong, you will need few sequences; if it is weak you will need more.
If you are analysing *several* alignments, each divided in two or more groups, you should normally have 25 or more alignments.

How should input data about the phenotypic group assignment of the sequences in the alignment(s) be entered into the program?
This key input must be hard-coded in the main program script of DeltaProt.

*What is the meaning of structural **level** in DeltaProt?*
By secondary structure we have these levels:
1= Alpha helices, 2: Beta sheets, C=Loops, 4= Unknown and 5= The whole molecule.
With level of 3D-stucture we mean:
1= Core, 2= Intermediate, 3=Surface, 4= Unknown and 5= The whole molecule.
If the secondary and/or 3D structure is known, or may be predicted, the analyses can be performed in each of these regions. Level is sometimes also called layers.

Why do DeltaProt rapport some p-values with negative sign?
The sign is just used to indicate the direction of an observed trend. A negative change is reported with a minus sign ahead, and a positive trend without any sign in front of the computed p-value.

Will DeltaProt be ported to other software platforms?
We have no present plan to make DeltaProt available on new platforms, but a R-version and a Python version may be interesting.

References

Adamczak, R., Porollo, A., Meller, J. (2004) *Accurate Prediction of Solvent Accessi-bility Using Neural Networks Based Regression.* Proteins: Structure, Function and Bioinformatics, 56: 753-767.

Benjamini, Y. and Yekutieli, D, (2001) *The control of the false discovery rate in multiple testing under dependency*. Ann Stat, 29 (4): 1165-1188.

Matlab (2005) The MathWorks, Inc. Natick, MA (US). www.mathworks.com

Thorvaldsen, S., Ytterstad, E. and Flå, T. (2006) *Property-dependent analysis of aligned proteins from two or more populations***.** Proceedings of the 4th Asia-Pacific Bioinformatics Conference (Eds.: T. Jiang et al.). Imperial College Press, pp. 169-178.

Thorvaldsen S, Hjerde E, Fenton C, et al. (2007) *Molecular characterization of cold adaptation based on ortholog protein sequences from Vibrionaceae species*. Extremophiles 11(5): 719-732

Thorvaldsen, S. and Ytterstad, E. (2009) *Environmental adaptation of proteins: Regression models with simple physicochemical properties*. Computational Biology and Chemistry. Vol. 33 (5): 351-356.