*Application Note*

# An Integrated System for Studying Residue Coevolution in Proteins

Kevin Y. Yip[1†], Prianka Patel[2†], Philip M. Kim[2], Donald M. Engelman[2], Drew McDermott[1] and Mark Gerstein[1, 2, 3*]

[1]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06511, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA

[3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

**ABSTRACT**

Residue coevolution has recently emerged as an important concept, especially in the context of protein structures. While a multitude of different functions for quantifying it have been proposed, not much is known about their relative strengths and weaknesses. Also, subtle algorithmic details have discouraged implementing and comparing them. We addressed this issue by developing an integrated online system that enables comparative analyses with a comprehensive set of commonly used scoring functions, including Statistical Coupling Analysis (SCA), Explicit Likelihood of Subset Variation (ELSC), mutual information, and correlation-based methods. A set of data preprocessing options are provided for improving the sensitivity and specificity of coevolution signal detection, including sequence weighting, residue grouping, and the filtering of sequences, sites and site pairs. A total of more than 100 scoring variations are available. The system also provides facilities for studying the relationship between coevolution scores and inter-residue distances from a crystal structure if provided, which may help in understanding protein structures.

**Availability:** The system is available at http://coevolution.gersteinlab.org. The source code and JavaDoc API can also be downloaded from the web site.

**Supplementary Information:** Additional materials can be found at http://coevolution.gersteinlab.org/coevolution/supp.jsp.

## 1 INTRODUCTION

Coevolution (covariation/correlated mutation) is the change of a biological object triggered by the change of a related object. For example, the coding genes of some interacting proteins are preserved or eliminated together in new species (Pellegrini et al., 1999), or have similar phylogenetic trees (Goh et al., 2000). At the amino acid level, some residues under physical or functional constraints exhibit correlated mutations (Suel et al., 2003, Gloor et al., 2005, Socolich et al., 2005). Coevolving residues in a protein are detected in a two-step process: 1) the multiple sequence alignment (MSA) of the protein and its homologs is constructed or obtained; 2) a coevolution score is calculated for each pair of sites in the MSA. There are two main difficulties in this process. First, a large number of scoring functions have been proposed in the literature (see Halperin et al., 2006 for a recent survey). It can be difficult to choose from them, as they exhibit subtle yet significant differences, and it is likely that different applications would require different functions. Second, coevolution analyses could be confounded by uneven sequence representations, insufficient evolutionary divergence, and the presence of gaps in the MSA. A successful coevolution study has to take all these details into account.

To address this need, we have developed an integrated system that provides a simple interface for preprocessing data, computing coevolution scores, and analyzing the results. It offers a great variety of scoring variations (over 100) for studying different types of proteins and testing different hypotheses. The workflow of the system is shown in Fig. 1. More details on the scoring functions, preprocessing options, and result analysis are provided below.

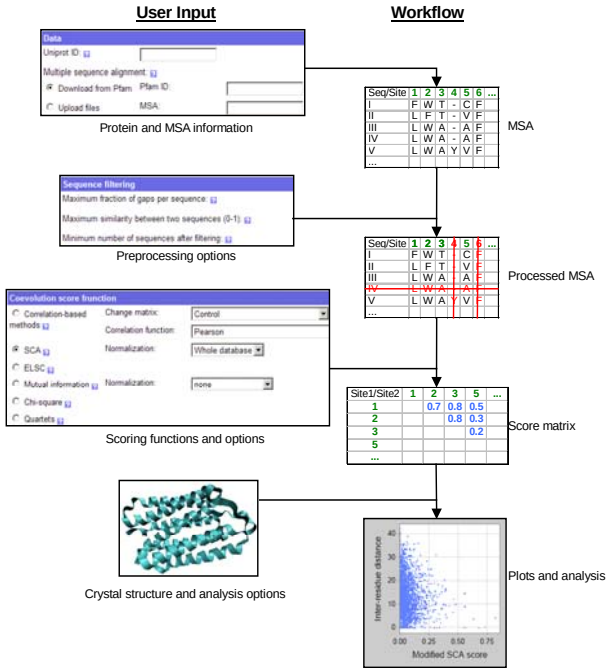


**Fig. 1.** The workflow of the system (a larger version can be found at the supplementary web site).

## 2 SCORING FUNCTIONS

### 2.1 Correlation-based functions

For a pair of sites i and j in an MSA, the correlation score (Gobel et al., 1994, Halperin et al., 2006) is computed as follows:

$$Cor(i,j) = \frac{2}{N(N-1)} \frac{\sum_{k<l} w_{kl} \left(s_{ikl} - \overline{s_i}\right)\left(s_{jkl} - \overline{s_j}\right)}{\sigma_i \sigma_j}$$

where $s_{ikl}$ is the score for substituting the $i$-th residue of sequence $k$ by that of sequence $l$, $\overline{s_i}$ and $\sigma_i$ are the mean and standard deviation of substitution scores at site $i$, $N$ is the number of sequences in the MSA, and $w_{kl}$ is the weight for the sequence pair $k$, $l$. If the two sites are coevolving in that radical substitutions at the first site are accompanied by radical substitutions at the second site, the correlation will be high. Our system provides the classical McLachlan matrix (McLachlan 1971) that scores substitutions based on the physiochemical properties of the residues, as well as matrices based on residue volume, pI, and hydropathy index, for studying the properties individually. Two variations are provided for each of them: the "absolute value version" considers only the magnitude, while the "raw version" also considers the direction of change, for detecting compensatory mutations. The correlation can be computed from raw values (Pearson correlation) or from value ranks (Spearman correlation, Pazos et al., 1997). Several schemes are provided for the weights $w_{kl}$, preventing false coevolution signals due to uneven sequence representation or site conservation.

### 2.2 Perturbation-based functions

[†]The authors have made equal contribution to the work

[*]To whom correspondence should be addressed.

The idea of perturbation-based functions is to perform a "perturbation" at a first site, and observe its effect on a second site. The Statistical Coupling Analysis (SCA) method (Lockless and Ranganathan 1999) defines a statistical energy term for a site, and computes the energy change at a second site when the first site is perturbed by retaining only the sequences with a certain residue.[1] The Explicit Likelihood of Subset Variation (ELSC) method (Dekker et al., 2004) is based on the same idea, but has the energy computations replaced by probabilities according to hypergeometric distributions. The mutual information (MI) method (Gloor et al., 2005) can be viewed as a generalized perturbation method that considers the subsetting of all twenty kinds of residues, and combines them by a weighted average according to their frequencies. To deal with finite sample size effects and phylogenetic influence, the normalization options in (Martin et al., 2005) are also provided.

### 2.3 Independence tests

The chi-square test (c.f. the OMES method, Larson et al., 2000) and the quartets method (Galitsky 2003) both identify site pairs that are unlikely to be independent. The former computes the p-value under the null hypothesis of independent sites. The latter counts the number of quartets in the two-dimensional histogram of residue frequencies that deviate considerably from the expectation.

## 3 PREPROCESSING OPTIONS

To improve the sensitivity and specificity of the functions, options are provided for preprocessing sequences, sites and site pairs.

### 3.1 Sequence filtering and weighting

Sequences that contain too many gapped positions or are too similar to others in the MSA (which might cause sites to appear coevolving) can be removed by specifying the gap and similarity thresholds respectively. A minimum number of sequences can also be specified to avoid small sample size effects.

A sequence weighting scheme based on the topology of the phylogenetic tree (Gerstein et al., 1994) and one based on Markov random walk are provided. Both schemes down-weigh sequences that are very similar to others in the MSA.

### 3.2 Site filtering

After sequence filtering, sites that contain too many gaps or are too conserved can be discarded. The former is likely non-informative, while the latter may artificially inflate some coevolution scores.

### 3.3 Site pair filtering

Sites that are close in the primary sequence may produce trivial coevolution signals that hide other more unexpected coevolution events. Such site pairs can be filtered by specifying the minimum sequence separation. It has also been observed that insertions/ deletions of multiple residues may create artificial coevolution signals (Patel et al., unpublished data). An option is provided for filtering site pairs that participate in the same gaps in too many sequences.

### 3.4 Other options

Grouping similar residues into a smaller alphabet may increase the sensitivity (Pollock et al., 1999). Our system provides two residue groupings proposed in the literature (Elcock and McCammon, 2001 and Guharoy and Chakrabarti, 2005). It has also been observed that gaps might give important coevolution signals (Patel et al., unpublished data). An option is provided for treating gaps as noise or as the 21$^{st}$ residue when computing coevolution scores.

## 4 SCORES ANALYSIS

In some proteins coevolving residues tend to be close to each other in the 3D structure (Dekker et al., 2004, Gloor et al., 2005). This suggests that the instability created by the mutation of a residue may be (partially) compensated for by a corresponding mutation of a close residue. Coevolution signals may thus convey some information about the protein structure. For instance it is interesting to study how well the coevolution scores predict the residue contact map (Halperin et al., 2006). Our system provides functions for plotting and analyzing the coevolution scores against inter-residue distances, and standard machine-learning techniques (e.g. ROC curve) for evaluating the effectiveness of the various coevolution functions in predicting interacting residues. A shuffling scheme for evaluating the significance of the scores is also provided in the program package for running locally.

[1] Our implementation provides an asymmetric SCA score matrix, as well as extra summarizing statistics. Details can be found at the supplementary web site.

## 5 EXAMPLE

We provide a worked example of our system in operation on the web site, which illustrates coevolution in the transmembrane protein bacteriorhodopsin due to physically constrained residues not adjacent in the primary sequence. The example can be easily loaded by clicking the corresponding link on the main page. Running the example will compute the coevolution scores between site pairs separated by at least 3 residues. The scatterplot for coevolution scores against inter-residue distances generated using a known PDB structure (Fig. 1) shows that residue pairs receiving high scores do tend to be closer in the crystal structure.

Due to the intensive computation involved in the score calculations, currently only one scoring function is allowed to be used each time. Anyone interested in performing large-scale comparisons can download the Java programs from the web site and run locally on most platforms (Windows, Macintosh, Linux, UNIX, etc.). Detailed installation instructions are provided on the web site.

## 6 DISCUSSION

Although the scatterplot in Figure 1, and other studies in the literature, have suggested some relationships between coevolution and physical constraints, to what extent could coevolution scores help understand physical structures remains unclear. We hope the current application could serve as a neutral tool for further exploration in this area.

The current system focuses on functions that do not assume any mutation models. Other functions, such as the likelihood method in (Pollock et al., 1999) and the Bayesian mutational mapping method (Dimmic et al., 2005) may be added in a later version.

Coevolution signals have been used in recent studies to predict sequence regions involved in protein-protein interactions with different levels of success (Pazos and Valencia 2002, Halperin et al., 2006). We plan on extending the system to include inter-protein residue coevolution in the next phase of development.

## ACKNOWLEDGMENT

We would like to thank Rama Ranganathan and William Russ for helpful discussions, and the anonymous reviewers for their valuable comments.

## REFERENCES

Dekker J. P. et al. (2004) A Perturbation-based Method for Calculating Explicit Likelihood of Evolutionary Co-variance in Multiple Sequence Alignments. *Bioinformatics* **20(10)** 1565-1572.

Dimmic M. W. et al. (2005) Detecting Coevolving Amino Acid Sites using Bayesian Mutational Mapping. *Bioinformatics* **21(suppl. 1)** i126-i135.

Elcock A. H. and McCammon J. A. (2001) Identification of Protein Oligomerization States by Analysis of Interface Conservation. *Proc. Natl Acad. Sci. USA* **98** 2990-2994.

Galitsky B. (2003) Revealing the Set of Mutually Correlated Positions for the Protein Families of Immunoglobulin Fold. *In Silico Biology* **3** 0022.

Gerstein M. et al. (1994) Volume Changes in Protein Evolution. *Journal of Molecular Biology* **236** 1067-1078.

Gloor G. B. et al. (2005) Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions. *Biochemistry* **44** 7156-7165.

Gobel U. et al. (1994) Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function, and Genetics* **18** 309-317.

Goh C. S. et al. (2000) Co-evolution of Proteins with their Interaction Partners. *Journal of Molecular Biology* **299(2)** 283-293.

Guharoy M. and Chakrabarti P. (2005) Conservation and Relative Importance of Residues across Protein-protein Interfaces. *Proc. Natl Acad. Sci. USA* **102** 15447-15452.

Halperin I. et al. (2006) Correlated Mutations: Advances and Limitations. A Study on Fusion Proteins and on the Cohesin-Dockerin Families. *Proteins: Structure, Function, and Bioinformatics* **63** 832-845.

Larson S. M. et al. (2000) Analysis of Covariation in an SH3 Domain Sequence Alignment: Applications in Tertiary Contact Prediction and the Design of Compensating Hydrophobic Core Substitutions. *Journal of Molecular Biology* **303** 433-446.

Lockless S. W. and Ranganathan R. (1999) Evolutionarily Conserved pathways of Energetic Connectivity in Protein Families. *Science* **286** 295-299.

McLachlan A. D. (1971) Tests for Comparing Related Amino-acid Sequences Cytochrome c and Cytochrome c551. *Journal of Molecular Biology* **61** 409-424.

Martin L. C. et al. (2005) Using Information Theory to Search for Co-evolving Residues in Proteins. *Bioinformatics* **21(22)** 4116-4124.

Pazos F. et al. (1997) Correlated Mutations Contain Information about Protein-Protein Interaction. *Journal of Molecular Biology* **271** 511-523.

Pazos F. and Valencia A. (2002) In Silico Two-hybrid System for the Selection of Physically Interacting Protein Pairs. *Proteins: Structure, Function, and Genetics* **47** 219-227.

Pellegrini M. et al. (1999) Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles. *Proc. Natl Acad. Sci. USA* **96** 4285-4288.

Pollock D. D. et al. (1999) Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure. *Journal of Molecular Biology* **287** 187-198.

Socolich M. et al. (2005) Evolutionary Information for Specifying a Protein Fold. *Nature* **437** 512-518.

Suel G. M. et al. (2003) Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nature Structural Biology* **10(1)** 59-69.