REVCOM: A Robust Bayesian Method for Evolutionary Rate Estimation

A.J. Bordner* and R.A. Abagyan
Molsoft LLC, 3366 North Torrey Pines Court, Suite 300, La Jolla, CA. 92037, U.S.A.

ABSTRACT
Motivation: Evolutionary conservation estimated from a multiple sequence alignment is a powerful indicator of the functional significance of a residue, and helps to predict active sites, ligand binding sites, and protein interaction interfaces. Many algorithms that calculate conservation work well provided an accurate and balanced alignment is used. However, such a strong dependence on the alignment makes the results highly variable. We attempted to improve the conservation prediction algorithm by making it more robust and less sensitive to (i) local alignment errors, (ii) overrepresentation of sequences in some branches, and (iii) occasional presence of unrelated sequences.

Results: A novel method is presented for robust constrained Bayesian estimation of evolutionary rates that avoids overfitting independent rates and satisfies the above requirements. The method is evaluated and compared with an entropy-based conservation measure on a set of protein interfaces. We demonstrate that about 60% of the analyzed protein interfaces are more conserved than the remaining surface at the 5% significance level. A consistent method to incorporate alignment reliability is proposed and demonstrated to reduce arbitrary variation of calculated rates upon inclusion of distantly related or unrelated sequences into the alignment.

Contact: bordner@molsoft.com

Supplementary Information: The protein-protein interface data set, multiple sequence alignments, and corresponding phylogenetic trees are available at http://www.molsoft.com/~bordner/REVCOM/

1 INTRODUCTION
The evolutionary conservation varies among amino acid sites due to differing degrees of functional constraints on them (Fitch and Margoliash, 1967; Uzzell and Corbin, 1971; Holmquist et al., 1983). Sites that are important for the protein’s tertiary structure and folding, enzymatic activity, ligand binding, or interaction with other proteins are generally more conserved. The greater conservation of residues in binding sites compared with other surface residues has been exploited by some methods that predict small ligand or protein-protein interfaces by mapping residue conservation values to the query protein surface (Landgraf et al., 2001; Lichtarge and Sowa, 2002; Pupko et al., 2002). The identification of conserved residues may be useful for identifying functionally important residues even in the absence of structural information. Binding site prediction results have applications in structure based drug design as well as protein functional assignment and suggest important residues for mutational analysis studies.

Models that account for the evolutionary relationship of the sequences through a phylogenetic tree should be less sensitive to the choice of sequences than methods based on residue frequencies in a corresponding multiple sequence alignment column. In particular, a large number of closely related sequences may give erroneously high residue conservation. Evolutionary tracing is one such method that utilizes a phylogenetic tree to identify residues that are identically conserved in a subtree (Lichtarge et al., 1996). The maximum tree depth at which a residue remains unchanged is used to rank the degree of conservation. This analysis was later modified to incorporate a quantitative model of residue substitutions (Landgraf et al., 1999). Another algorithm, ConSurf, used a maximum parsimony tree to calculate a site conservation score as the number of substitutions weighted by their physicochemical distance (Armon et al., 2001). A study by Pupko et al. (Pupko et al., 2002) describes a method for a maximum likelihood method estimation of independent evolutionary rates assuming a homogeneous Markov model of residue substitution. The maximum likelihood method (Neyman, 1971; Felsenstein, 1981) is not subject to errors present in the maximum parsimony principle including

*To whom correspondence should be addressed
A.J. Bordner\textsuperscript{1} and R.A. Abagyan

systematic overestimation of conservation due to neglect of backward and parallel substitutions and no dependence on branch lengths. It also has the advantage that a fast recursive algorithm may be used to calculate likelihoods for a given tree (Felsenstein, 1981).

However straightforward estimation of independent site rates suffers from overfitting since there are as many parameters as sites (Felsenstein, 2001). One solution to this problem is to assume that the rates come from a probability distribution, usually assumed to be a gamma distribution. Yang used a discrete approximation to the gamma distribution to incorporate rate variation in phylogeny determination using the maximum likelihood method (Yang, 1993, 1994). Although there is no evolutionary process that supports it, a gamma rates distribution is often used since it is defined on the correct interval (positive rates). It also leads to a negative binomial distribution for total substitution frequencies in a simple Poisson model of site substitution probabilities. An early paper by Uzzell and Corbin (Uzzell and Corbin, 1971) supported a gamma rates distribution by showing that the negative binomial distribution fits sequence data better than a Poisson distribution. Because of this evidence and mathematical convenience we also use the gamma distribution, but as a Bayesian prior rates distribution.

A lack of robustness, or excessive sensitivity of a conservation calculation to the input multiple sequence alignment is a general problem for the existing methods. This sensitivity may be due to overrepresentation of a particular subfamily or alignment errors. While distantly related sequences have the potential to provide strong evidence of residue conservation they also generally introduce more noise. Three source of error at large evolutionary distances may affect the site rate calculation: (1) local alignment errors due to ambiguities in divergent sequence segments (2) alignment to a nonhomologous sequence mistakenly picked out by a database search, and (3) uncertainties in the residue substitution matrix due to alignment errors or extrapolation to large distances. All of these errors are expected to become more likely at larger distances and are neglected in most conservation prediction methods. This is important, particularly for automatically generate alignments which may include nonhomologous sequences or alignment errors for distantly related sequences. While iterative alignment methods, such as PSI-BLAST, have high sensitivity in detecting remote homologues (Park et al., 1998) they are especially susceptible to including nonhomologous sequences due to overly permissive parameters or profile drift (Sjölander, 2004). We introduce the Robust EVolutionary COnservation Measure (REVCOM) which consistently incorporates the alignment reliability into the likelihood calculation in order to render the method more robust to the inclusion of distantly related sequences. The resulting likelihood for each site may be interpreted as a sum of likelihoods for all possible trees with a subset of sequences removed, weighted by the probability that the excluded sequences are unreliable and the included ones are reliable.

A large nonredundant data set of 1568 protein-protein interfaces with available X-ray crystal structures was also compiled in order to test the conservation algorithm. Since they are larger in general than ligand binding interfaces the variance in conservation statistics for protein-protein interfaces is expected to be lower. We compared the REVCOM method with a simple entropy base method that does not use evolutionary trees. Both the ability of the methods to detect the higher conservation in protein-protein interfaces and their stability with the inclusion of distantly related sequences in the alignment were evaluated.

2 SYSTEMS AND METHODS

2.1 Residue substitution model

Amino acid substitutions at sites are assumed to follow independent time-homogeneous Markov processes with the Jones-Taylor-Thorton (JTT) matrix used to calculate substitution probabilities (Jones et al., 1992). The substitution matrix for a given evolutionary distance was extrapolated from the matrix given in this reference for a distance corresponding to one percent accepted point mutation, or one PAM.

2.2 Multiple sequence alignments

Multiple sequence alignments were generated automatically for each protein sequence in the data set. First the BLAST program (Altschul et al., 1990) with an E-value cutoff of 0.1 was used to collect similar sequences from the NCBI nr database. Sequences with > 90\% identity to another sequence in the set were then iteratively removed. The ClustalW program (Thompson et al., 1994) with default alignment parameters (Gonnet 250 scoring matrix with gap opening = 10 and gap extension = 0.1) was used to align the remaining sequences.

2.3 Phylogenetic trees

Phylogenetic trees were then generated for each alignment using the neighbor-joining algorithm (Saitou and Nei, 1987) as implemented in the Quicktree (Howe et al., 2002) program. A weighted least squares fit using pairwise PAM distances between sequences was then used to recalculate the branch lengths. The PAM distance $t$ was calculated by inverting the expression for the expected fraction of identical residues $q(t)$

$$q(t) = \sum_{i=1}^{\infty} (M^t)_{ii} f_i$$

where $M^t = \exp(t \log M)$ is the JTT matrix for distance $t$, defined by the matrix exponential, and $f_i$ is the occurrence probability of amino acid type $i$. The variance $\sigma_i^2$ in the distance was approximated from the binomial variance

\[\sigma_i^2 = \frac{\frac{1}{10} \log M_{ii} f_i}{f_i (1-f_i)}\]
using the delta method

$$\sigma^2_t \approx \left( \frac{dq(t)}{dt} \right)^{-2} \frac{q(1-q)}{n}. \quad (2)$$

The weighted least squares fit then minimizes the sum of squares of weighted residuals

$$SS = \sum_{i=1}^{N(N-1)/2} \frac{(d_i - \sum_{j=1}^{N-2} C_{ij} b_j)^2}{\sigma^2_i} \quad (3)$$

where $N$ is the number of sequences, $d_i$ is the pairwise PAM distance, $b_j$ is the branch length and $C$ is a matrix with elements $C_{ij} = 1$ when branch $j$ is in the unique path joining sequences in the pair $i$ and 0 otherwise (Bulmer, 1991). Defining a diagonal variance matrix $V_{ii} = \sigma^2_i$ the estimates of the branch lengths are $b = (C^T V^{-1} C)^{-1} C^T V^{-1} d$.

### 2.4 Simple Residue Conservation Model

We compared our rate prediction model with a simple entropy-based model in which the site conservation is calculated for each alignment column using the Shannon entropy

$$S = -\sum_{i=1}^{20} f_i \log f_i \quad (4)$$

where $f_i$ are the residue occurrence frequencies for the column (Valdar, 2002; Mihalek et al., 2004). Gaps are not included in the calculation of the residue frequencies.

### 2.5 Protein-protein interface data sets

A data set of protein intermolecular interfaces in complexes was compiled from biological unit information in the Protein Data Bank (PDB) (Berman et al., 2000) archive using the ICM scripting language (Molsoft, LLC, 2004). Only the first biological unit was included and mmCIF format files were utilized since they contain biological unit information for all entries. Only X-ray structures of proteins with $\geq 20$ residues were considered and pairs of interacting proteins were clustered at 30% sequence identity. Alignments of the protein sequences in each cluster to a representative sequence were then used to compare interface residues. Two interfaces on a protein were considered distinct if their residue sets, referred to the representative sequence, overlapped less than 20%. The highest quality structure, with the least missing coordinates and the highest resolution, for each unique protein-protein interface was then included in the data set. Interfaces including immune system proteins that are highly polymorphic proteins or undergo somatic mutation, namely MHC, T-cell receptors, and antibodies, were excluded. Finally, only interfaces containing at least 10 residues were included in the set because of the large variance of the residue-based statistics as well as difficulties in validating the interface prediction for smaller interfaces. Since the protein-protein interfaces are considered as surface regions on individual proteins, both orderings of the non-identical interacting proteins were considered distinct. The resulting set had 1568 protein-protein interfaces (1401 unordered pairs), of which 549 were homodimers, 131 were heterodimers, and the remaining 888 were multimers.

### 2.6 Statistical Tests

The Mann-Whitney-Wilcoxon rank sum test was used to compare the evolutionary rates for the interface and non-interface regions. This nonparametric test was used since the underlying distribution of the posterior rates is unknown and non-normal.

### 3 ALGORITHM

#### 3.1 Evolutionary site rates estimation

The REVCOM method calculates evolutionary rates for sites in a reference protein sequence using the following steps:

- Perform a BLAST search of the sequence database using the reference sequence as a query
- Remove redundant sequences (>90% identity)
- Construct a multiple alignment of all sequences
- Calculate all pairwise PAM distances between sequences from fraction identity using Eq. 1
- Calculate phylogenetic tree from pairwise distances using neighbor-joining algorithm
- Recalculate branch lengths using weighted least squares method
- Find maximum likelihood estimate of $\alpha$ parameter in prior gamma distribution
- Calculate site rates by averaging over posterior rates distribution

P-values from the BLAST search in the first step are used as estimates of alignment reliability in the final two steps.

#### 3.2 Maximum Likelihood Estimate of Gamma Distribution Parameter

The likelihood that is maximized for the $\alpha$ parameter estimate is

$$L(\alpha) = \prod_m \int_0^\infty p(\bar{X}_m | r) f(\alpha, r) dr \quad (5)$$

where $f(\alpha, r)$ is the gamma distribution with mean 1

$$f(\alpha, r) = \frac{\alpha^a r^{a-1} e^{-\alpha r}}{\Gamma(\alpha)} \quad (6)$$

and $p(\bar{X}_m | r)$ is the likelihood for site $m$, given the rate $r$, $\bar{X}_m$ is a vector of amino acid types in column $m$ of the multiple sequence alignment. Nonstandard amino acid symbols, e.g. $X$, and gaps are excluded from the calculation by removing the corresponding branches in the tree for that site.

3
3.3 Accounting for Alignment Reliability

One new feature in the REVCOM method is the use of an alignment reliability probability in order to prevent distantly related or nonhomologous sequences from inordinately affecting the likelihood in Eq. 5 and consequently the predicted rates. The probabilities of each sequence to be misaligned with a reference sequence, for which the rates are to be calculated, are used. The misalignment probabilities for different sequences will be assumed to be independent. This allows them to be easily incorporated into the maximum likelihood calculation. Furthermore, the misalignment probabilities will be chosen to be constant for all alignment columns and equal to the BLAST p-values, \( 1 - \exp(-E_i) \), with \( E_i \) the E-value from the database search with the reference sequence of the protein onto which the rates will be projected. This only explicitly accounts for error (2) mentioned above, however this is a serious global alignment error whose statistical distribution is known (Altschul and Gish, 1996). Local, or column dependent, alignment reliability measures (Vingron and Argos, 1990; Chao et al., 1993; Mevissen and Vingron, 1996; Abagyan and Batalov, 1997; Schlosshauer and Ohlsson, 2002; Löyttyniemi and Milinkovitch, 2003) may also be used.

The site likelihood \( p(\tilde{x}_m|\tilde{r}) \) is calculated recursively using a modification of the method of Felsenstein (Felsenstein, 1981).

\[
p(\tilde{x}_m|\tilde{r}) = \sum_{j=1}^{20} f_j L_i(0) \tag{7}
\]

where \( f_j \) is the residue probability, as above, and \( L_i(I) \) is the conditional likelihood for a subtree based at node \( I \) with \( I = 0 \) the root node. The following recursion relations are then used to calculate \( L_i(0) \) (refer to Fig. 1):

\[
L_i(I) = \sum_{j=1}^{20} \sum_{k=1}^{20} [(1 - q_M)(\text{M}^r b_M)_{ij} L_j(J) + q_M]
\]

\[
\times [(1 - q_N)(\text{M}^r b_N)_{ik} L_k(K) + q_N]
\]

\[
L_i(I) = \delta_{j,s_i}, \quad I \in \text{leaf nodes}. \tag{8}
\]

\( b_M \) is the length and \( q_M \) the misalignment probability of branch \( M \) and \( S_j \) is the residue type in the sequence associated with leaf node \( I \) at the site. \( q_M \) for a branch connected to a leaf node is equal to the the misalignment probability of the corresponding sequence and is equal to zero for all other branches. The total site likelihood is then the sum of the likelihoods for the site residues given phylogenetic trees with all combinations of the non-reference sequences removed, weighted by the probability that the missing sequences are misaligned and the included sequences are correctly aligned. For example, the four terms that contribute to the likelihood sum and their respective weights are shown in Fig. 2 for the case of one reference sequence \( R \) and two other sequences \( M \) and \( N \).

3.4 Bayesian Rates Estimate

The site rate distribution with the estimated parameter \( \hat{\alpha} \) may then be used as a prior distribution to calculate the posterior rate distribution using Bayes’ formula

\[
p(r|\tilde{x}_m) = \frac{p(\tilde{x}_m|r) f(\hat{\alpha}, r)}{p(\tilde{x}_m)} \tag{9}
\]

with

\[
p(\tilde{x}_m) = \int_0^{\infty} p(r|\tilde{x}_m) dr. \tag{10}
\]

A Bayesian estimate of the site rate, \( \hat{r} \), is then calculated as the average of this posterior distribution

\[
\hat{r} = \int_0^{\infty} p(r|\tilde{x}_m) r dr. \tag{11}
\]

Using the average posterior distribution as an estimate corresponds to minimizing a square-error loss function (see e.g. Hogg and Craig, 1978).

4 IMPLEMENTATION

Several numerical approximations were used to speed up the calculation. A conjugate gradient iterative method was used to solve the linear system of equations resulting from the least squares fit of PAM distances. Also components of the JTT substitution matrix, which are continuous functions of the evolutionary PAM distance, were approximated by Chebyshev polynomials for fast calculation (Pupko and Graur, 2002). Finally, the integral in Eq. 5 for the total likelihood may be efficiently calculated using Gaussian quadrature based on Laguerre polynomials (Press et al., 1992), as discussed in Felsenstein, 2001.

5 DISCUSSION

5.1 Comparison with Entropy Based Conservation Measure

The site rates calculated using the REVCOM method were compared with the simple entropy conservation measure by examining their ability to detect statistically significant differences in evolutionary conservation between protein-protein interface residues and non-interface surface residues. Statistical tests with an alternative hypothesis of higher residue conservation in the interface were performed at three different significance levels, 5 %, 1 %, and 0.1 % using both conservation measures. All interfaces in the data set were analyzed. The number of proteins with significantly higher conservation in their interaction interfaces as well as the median p-values are given in Table 1 for each method. It is apparent that the REVCOM method discriminates conservation differences better than the entropy measure since it detects significant conservation differences in more interfaces at all significance levels. Overall approximately 60 % of the interfaces have significantly higher residue conservation than the non-interface surface residues.
5.2 Robustness to Inclusion of Distantly Related Sequences

Site rate predictions may be compared with predictions using alignments containing additional distantly related or nonhomologous sequences in order to verify that including misalignment probabilities renders the method more robust. First, the NCBI nr database was searched with two different BLAST E-value cutoffs, 0.1 and 10.0, using sequences for each of the 1568 proteins in the data set as queries. Next, rates were calculated for the 805 proteins for which additional sequences were found at the higher E-value cutoff. The root mean square deviation in rates was only 0.16 for the REVCOM method as compared with 0.25 for the same method without misalignment probabilities. A histogram of the differences in the rates as well as in the entropy measures with an expansion of the multiple alignment is shown in Figure 3. The greater robustness from accounting for alignment reliability is apparent from the histogram of the REVCOM method, which is more strongly peaked about zero rate difference. The histogram also shows that expansion of the multiple alignment causes a large systematic increase in the entropy conservation measure whereas the differences for the REVCOM method are more symmetrical about zero.

Next, we illustrate how the misalignment probabilities render the method more robust to the inclusion of nonhomologous sequences by a specific example. Site rates are calculated for the SH2 domain using a reliable alignment and an alignment with additional nonhomologous sequences and compared. The Pfam seed alignment (Bateman et al., 2002) with 58 sequences is used for the reliable alignment and 23 (≈ 40%) random nonhomologous sequences are then added and aligned to a profile of the original sequences using ClustalW (Thompson et al., 1994) for the test alignment. The median site rate differences between the reliable and test alignments is 0.33 without including alignment reliability and only 0.065 when it is included. This is due to the decoupling of the nonhomologous sequences in the site likelihood calculation when misalignment probabilities are accounted for.

5.3 Conservation in Protein-Protein Interfaces

It is important to compare conservation of interface residues with that of other surface residues rather than with all protein residues. This is because the lower conservation of surface residues compared with buried residues yields a greater contrast. In fact, one study compared the evolutionary conservation of protein-protein interface residues with the entire protein sequence and found little difference (Grishin and Phillips, 1994). However, other studies, which compared interface residues with surface residues, found significant differences in conservation (Valdar and Thornton, 2001; Caffrey et al., 2004). Our results corroborate this observation but with a much larger data set than those used in these earlier papers.

The site rates may be displayed on the protein surface to identify protein-protein interaction interfaces and active sites. An example of E. coli malate dehydrogenase is shown in Figure 4. This homodimeric enzyme uses NAD as a cofactor to reversibly oxidize malate to oxaloacetate. It is an essential component of the citric acid cycle with orthologues in both prokaryotes and eukaryotes. The figure shows that residues in a surface region extending from the dimer interface to the cofactor and substrate binding sites have lower rates, i.e., higher conservation, in general than the remaining surface residues. The lower evolutionary rates for the interface is also evident from the low p-value of $5.2 \times 10^{-8}$ for the statistic comparing interface and non-interface residue rates.

Although the majority of protein-protein interfaces in our data set are more conserved than the remainder of the surface, the paper of Caffrey et al. (Caffrey et al., 2004) concluded that conservation alone is insufficient to predict the interfaces when the surface is divided into small patches. Extending their analysis to a larger data set or using different conservation measures, such as the one presented here, may yield better results. However, including physiochemical, geometric or residue distribution properties, such as those considered by Jones and Thornton (Jones and Thornton, 1997) in their patch analysis of protein-protein interfaces, can only improve the prediction performance. Thus an accurate protein-protein interface prediction method should account for residue conservation but also include other discriminating properties.

5.4 Future Directions

There are several possible extensions of the REVCOM method presented here. A local, or column dependent, alignment reliability measure (Vingron and Argos, 1990; Chao et al., 1993; Mevissen and Vingron, 1996; Abagyan and Batalov, 1997; Schlosshauer and Ohlsson, 2002; Löytynoja and Milinkovitch, 2003) may be used for calculating the misalignment probabilities instead of the global measure. However, the most appropriate measure, not described in these references, would be the probability of a local misalignment error between a given sequence and the reference sequence in the context of the multiple sequence alignment. Since misalignment probabilities are required by the method described in this paper a local alignment reliability index must be converted into a probability, e.g., by calibrating it against a structural alignment database (Mevissen and Vingron, 1996). It is also interesting to investigate the use of different prior rates distributions. The distribution should not have many parameters because estimation of the gamma distribution parameter is currently the most computationally intensive step in the rates calculation. Finally, since there is evidence that alignment gaps are less frequent in protein-protein interfaces, at least for permanent dimers (Caffrey et al., 2004), accounting for
gaps in the stochastic evolutionary process, rather than treating them as missing data, may result in more accurate site rates.

It would be informative to divide the interface data set into permanent and transient protein-protein interactions to investigate the differences in interface residue conservation. Caffrey et al., 2004 found that central interface residues were more conserved than peripheral interface residues in permanent interfaces but not transient ones, but their data set was considerably smaller than the one used here. The difficulty is in automatically assigning binding affinities to a large number of protein interactions. Another issue that has not been thoroughly investigated is whether residues that contribute the most to the binding free energy are more conserved than other interface residues. Previous studies have shown, using alanine-scanning mutagenesis results, that a small number of interface residues contribute to the majority of the binding energy (Clackson and Wells, 1995; Bogan and Thorn, 1998). One paper (Hu et al., 2000) identified polar residues that are conserved in a structural alignment of protein-protein interfaces and speculated, based on the agreement with alanine scanning data, that these may be “hot spot” residues that strongly influence binding.

6 CONCLUSION

We have presented a method, REVCOM, to calculate residue conservation that accounts for the evolutionary relationships of the sequences and avoids overfitting by using a Bayesian framework. This method also accounts for alignment errors by summing the likelihoods of truncated phylogenetic trees which are weighted by the probability that the missing sequences were unreliable and the included ones were reliable. The resulting conservation measure was shown to discriminate conservation differences for protein-protein interfaces better than a column entropy measure and to prevent distantly related sequences from overly affecting the rates. The evolutionary rates provided by the REVCOM method are useful for identifying surface residues in enzyme active sites and small ligand or protein binding sites.

ACKNOWLEDGEMENTS

This work was funded by a grant from the Department of Energy (No. DE-FG03-01ER83282).

REFERENCES

Proteins, 39, 331–342.


<table>
<thead>
<tr>
<th>Conservation Model</th>
<th>p-value cutoff</th>
<th>median p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Rates</td>
<td>962 770 567</td>
<td>0.0113</td>
</tr>
<tr>
<td>Entropy</td>
<td>831 617 407</td>
<td>0.0352</td>
</tr>
</tbody>
</table>

Table 1. Number of protein-protein interfaces in the set of 1568 that have greater evolutionary conservation as compared with non-interface surface residues at the given significance levels. The median p-values for the complete set are also given.
Fig. 1. Tree branch and vertex labels for Eq. 8
Fig. 2. Schematic representation of the four terms contributing to the total likelihood for the case of one representative sequence $R$ and two other sequence $M$ and $N$. 

\[(1-q_{M})(1-q_{N})\]  
\[+ q_{M}(1-q_{N})\]  
\[+ (1-q_{M})q_{N}\]  
\[+ q_{M}q_{N}\]
Fig. 3. Comparison of the robustness of three conservation calculation methods. Histograms of conservation differences upon multiple alignment expansions are shown. A perfectly robust conservation measure would have all differences identically zero. Since the area under each curve is equal, a higher peak at zero difference indicates higher robustness.
Fig. 4. Surface residue site rates for *E. coli* Malate Dehydrogenase homodimer (PDB entry 1EMD (Hall and Banaszak, 1993)) calculated using REVCOM method. The surface of one protein is colored by rates with blue, white, and red corresponding to low, intermediate, and high rates respectively. The most conserved residues are clustered near the dimer interface, NAD cofactor (shown in stick representation), and substrate (not visible beneath surface to left of cofactor).