J Mol Evol (1996) 42:313–320

JOURNAL OF MOLECULAR EVOLUTION

Probabilistic Reconstruction of Ancestral Protein Sequences

Jeffrey M. Koshi,[1] Richard A. Goldstein[1,2]

[1] Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055, USA
[2] Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055, USA

Received: 6 March 1995 / Accepted: 5 October 1995

Abstract. Using a maximum-likelihood formalism, we have developed a method with which to reconstruct the sequences of ancestral proteins. Our approach allows the calculation of not only the most probable ancestral sequence but also of the probability of any amino acid at any given node in the evolutionary tree. Because we consider evolution on the amino acid level, we are better able to include effects of evolutionary pressure and take advantage of structural information about the protein through the use of mutation matrices that depend on secondary structure and surface accessibility. The computational complexity of this method scales linearly with the number of homologous proteins used to reconstruct the ancestral sequence.

Key words: Bayesian statistics — Evolutionary reconstruction — Homologous sequences — Protein evolution — Maximum likelihood

Introduction

The proteins that exist at our current stage of evolution represent a minuscule subset of the proteins that have existed since life began. Examination of the ancestors to modern-day proteins would undoubtably give great insight into the properties of current proteins as well as into the process of evolution. If the sequence of an ancestral protein can be re-created, the protein can be synthesized and expressed, and its characteristics can be determined experimentally (Malcolm et al. 1990; Stackhouse et al. 1990; Shih et al. 1993).

Barring the presence of historically preserved DNA (Higuchi et al. 1984; Paabo 1989; Cooper et al. 1992; DeSalle et al. 1992), these ancestral amino acid sequences must be reconstructed based on the known sequences of the current descendents. Generally, this recreation has been done using some variation on the methods of maximum parsimony (MP) (Fitch 1971; Moore et al. 1973; Holmquist 1979; Czelusniak et al. 1990) or maximum likelihood (ML) (Felsenstein 1981; Saitou 1990; Yang 1994). Both of these approaches involve a step-wise construction of an evolutionary scenario that is considered optimal in either minimizing total mutational steps (MP) or in maximizing the likelihood of the mutations occurring (ML). The MP method, as well as the closely related inferential method (Libertini and Donato 1994), generally not only discards everything except the maximally parsimonious schemes, but also cannot distinguish between scenarios of equal parsimony. In contrast, approaches based on the ML methodology can not only generate the most likely ancestral sequence but can also consider all other possibilities and compute their respective probabilities. These suboptimal reconstructed sequences are important in that they can provide information about other possibilities, suggesting alternatives that can be tested using biochemical means.

Much of the reconstruction work up to this point has

Abbreviations: ML: maximum likelihood; MP: maximum parsimony; PAM: point-accepted mutations
Correspondence to: R.A. Goldstein

been based on nucleic acid substitutions, and amino acid mutations are subsequently reconstructed from the nucleic acids. Benner and co-workers have shown that for short evolutionary intervals mutations can be understood using nucleic acid substitution rates. For longer intervals, however, it is the requirements that biology places on protein function and structure that constrain the evolutionary path (Benner et al. 1994b). By considering the evolutionary pathway on the amino acid level it is possible to take advantage of these constraints in order to more correctly model the longer-scale molecular evolution process. This approach also allows us to directly include information about the local structure at every site in the reconstruction by using a priori probabilities and mutation matrices specific for each type of local structure.

In previous work, we have constructed the structure-dependent mutation matrices mentioned above. They are based on phylogenetic models of protein evolution and represent the probability of mutation from one residue to another in a given period of evolutionary time (Koshi and Goldstein 1995). This explicit modeling of the evolutionary history gives us the tools needed to do a probabilistic reconstruction of ancestral sequences at the amino acid level. Because we are using an ML approach based on Bayesian statistics, we can determine the probability of any given amino acid existing at any stage of evolution as well as the probability of any particular evolutionary path. We can re-create the most likely ancestral sequence and also evaluate the relative probability of any other sequence existing at that point in evolutionary time. As we have developed specific mutation matrices for various combinations of secondary structure and surface accessibility, we can use structural information about the proteins to assist in this reconstruction.

In this paper, we explore this methodological framework for evolutionary reconstruction, first demonstrating how the accuracy of this approach depends upon the number of homolog proteins available and their evolutionary distance. Finally we apply this approach to the reconstruction of the ancestral sequence of the ribonuclease superfamily.

Methodology

Our approach to the construction of ancestral sequences follows our original derivation of optimized mutation matrices (Koshi and Goldstein 1995). We start with a set of homologous proteins connected by a known evolutionary tree *T,* and the amino acids found at a given location in the previously aligned sequences of the homologous proteins, $\{A_i\}'$, where the prime indicates that our knowledge is restricted to proteins at our current stage of evolutionary history. In addition, we assume we have a reasonable approximation to the mutation matrix *M* providing the probability of mutation from one residue to another. *M* can be a function of the secondary structure or surface accessibility at that location in the protein, or it can include any other available information. We are interested in computing $P(A_r|\{A_i\}',M,T)$, the conditional probability of a given amino acid A_r at the root location, given $\{A_i\}'$, *M,* and *T.* This probability can be easily computed using Bayes' Theorem:

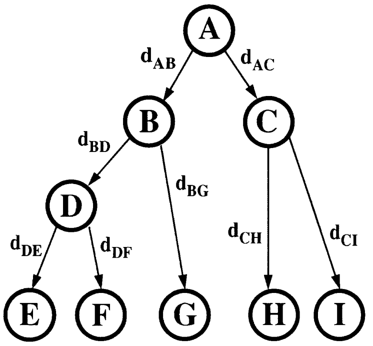

Fig. 1. Representation of an evolutionary tree corresponding to a particular set of aligned homologous proteins. The nodes *E, F, G, H,* and *I* represent amino acids in present-day proteins. The composition of nodes *A, B, C,* and *D,* which represent proteins in the evolutionary past, are unknown. d_{XY} represents the evolutionary distance between nodes *X* and *Y.*

$$P(A_r\,|\,\{A_i\}',\,M,\,T) = \frac{P(\{A_i\}'\,|\,A_r,\,M,\,T)P(A_r)}{P(\{A_i\}'\,|\,M,\,T)} \tag{1}$$

where $P(\{A_i\}'A_r,M,T)$ is the conditional probability of observing the particular set of present-day amino acids for a given mutation matrix, evolutionary tree, and root amino acid A_r, and $P(A_r)$ is the a priori probability of a given root amino acid A_r, obviously independent of the tree *T* and mutation matrix *M.* $P(\{A_i\}'M,T)$ simply serves to normalize the probabilities so the sum of $P(A_r\{A_i\}',M,T)$ over all possible values of A_r equals 1.

$P(A_r)$ can be approximated by considering the relative probability of various amino acids in current proteins, including any dependence on the location of the residues in the protein. The calculation of $P(\{A_i\}'A_r,M,T)$ is more involved. Consider a simple example where the evolutionary tree has the structure shown in Fig. 1. While we are calculating a value of this probability for an amino acid A_A at root node *A,* we still do not know the identity of the residues at nodes *B, C,* and *D.* We must, therefore, sum over all 21 possibilities (i.e., all 20 amino acids, plus gaps) at each of these nodes. $P(\{A_i\}'A_A,M,T)$ is then given by the probability of the mutations necessary to produce the amino acids at nodes *E* through *I,* summed over all possible pathways from the root amino acid to these leaves.

$$P(A_E,\,A_F,\,A_G,\,A_H,\,A_I\,|\,A_A,\,M,\,T) = \sum_{A_B,A_C,A_D} T_{A_A\rightarrow A_B}T_{A_B\rightarrow A_D}T_{A_D\rightarrow A_E}T_{A_D\rightarrow A_F}T_{A_B\rightarrow A_G}T_{A_A\rightarrow A_C}T_{A_C\rightarrow A_H}T_{A_C\rightarrow A_I} \tag{2}$$

where $T_{A_A\rightarrow A_B}$ represents the probability of amino acid A_A mutating to amino acid A_B, which can be computed with knowledge of the mutation matrix *M* and the evolutionary distance between the various branching points in the phylogenetic tree.

These relationships can easily be generalized to more complex evolutionary tree structures, allowing us to calculate $P(\{A_i\}'A_A,M,T)$ and thus $P(A_A\{A_i\}',M,T)$, for any specific tree structure, leaf composition, and mutation matrix. This method is easily generalizable to re-create the amino acids at other locations in the phylogenetic tree besides the root. For instance, in the example in Fig. 1, the identity of residues A_H and A_I can provide information about the probability of a given residue having existed at ancestral node *B* by influencing the probability distribution of residues at node *A.*

The calculation starts with aligned sequences and a reconstructed evolutionary tree. This can be produced with software already available and can take advantage of any other information about the proteins or organisms concerned. In this paper, we use the alignments and phylogenetic trees produced by ClustalV (Higgins et al. 1992). We also use the optimized mutation matrices described previously (Koshi and Goldstein 1995). Assuming a constant mutation rate and no double mutations at any location in the sequences between consecutive nodes, $T_{A_A \rightarrow A_B}$, the probability of mutation from residue A_A to A_B in evolutionary time d_{AB}, is given by:

$$T_{A_A \rightarrow A_B} = M(A_A, A_B) d_{AB} | A_A \neq A_B \qquad (3)$$

and

$$T_{A_A \rightarrow A_A} = 1 - d_{AB}(1 - M(A_A, A_A)) \qquad (4)$$

where $M(A_A, A_B)$ is the corresponding element in the desired mutation matrix M and

$$M(A_A, A_A) = 1 - \sum_{A_x \neq A_A} M(A_A, A_X) \qquad (5)$$

The assumption of no double mutations between consecutive nodes does not imply a total lack of multiple mutations during the evolutionary process. In fact, as mentioned, *all* possible evolutionary trajectories are explicitly considered. This assumption of lack of multiple mutations is not an essential part of this method; because the evolutionary distances are computed externally, it would be straightforward to use a more complicated model for the mutation rates. This would involve raising the mutation matrix to a power proportional to the time between nodes and using the resulting matrix in place of the $M(A_A, A_B)$ values seen in the above equations (Dayhoff and Eck 1968). However, as all of the evolutionary paths considered here were relatively short, this was not done.

This whole approach is based on treating each location in the set of aligned sequences independently. In particular, we treat gaps the same as any amino acid, where the probability of a gap ‘‘mutating’’ to any other amino acid corresponds to the probability of an insertion of that amino acid into the sequence. It would be possible to construct a more realistic model of insertions and deletions incorporating the cooperative nature of these events.

The number of paths considered rises exponentially with the number of homologs that are being used to reconstruct the path. In practice, however, the calculation is much simpler. As noted by Felsenstein, a binary tree can be evaluated in linear time by traversing it from leaves to root (Felsenstein 1973, 1981). More specifically, every node in a binary tree has two filial nodes below and directly connected to the paternal node. For instance, in Fig. 1, nodes D and G are filial to paternal node B. For every paternal node k with filial nodes m and n,

$$P(\{A_i\}''_k | A_k, M, T)$$

$$= \sum_{A_m, A_n} T_{A_k \rightarrow A_m} P(\{A_i\}''_m | A_m, M, T)\, T_{A_k \rightarrow A_n} P(\{A_i\}''_n | A_n, M, T) \qquad (6)$$

where $\{A_i\}''_k$ is the set of amino acids in homologous proteins at our current stage of evolutionary history that are direct descendents of node k. The calculation of $P(\{A_i\}''_k A_k, M, T)$ for every possible A_k involves at most 21^3 calculations, assuming that the values of $P(\{A_i\}''_m A_m, M, T)$ for the filial nodes have already been calculated. As the number of nodes is one fewer than the number of homologs, the calculational complexity varies only linearly with the number of sequences being considered.

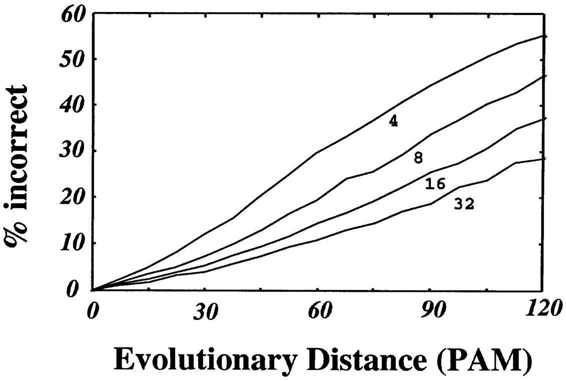

Fig. 2. Ten 500-residue random ancestral sequences were generated and allowed to mutate; the ancestral sequence was then re-created based on the sequences of these mutated homologs. Percentage incorrect is plotted vs evolutionary distance from the top to the bottom of the tree (equal steps between each node) for various numbers of homologs.

Results

We initially investigated two issues: how the accuracy of the reconstruction depends on the evolutionary distances involved and how many homologs are necessary for an accurate reconstruction. In order to explore these questions, we prepared artificial data sets by modeling the mutation process. We started with 10 model ancestral proteins 500 residues long whose secondary structure and surface accessibility represented the overall distribution seen in real proteins (15.2% inside helix, 13.4% outside helix, 14.4% inside sheet, 7.5% outside sheet, 7.6% inside turn, 18.7% outside turn, 13.3% inside coil, 9.9% outside coil). The residues at each position were selected based on the relative propensity for each of the amino acids to be in such a location. We then modeled evolution and speciation, allowing the evolutionary tree to branch periodically and symmetrically. This resulted in 2^n present-day homologs, with a fixed interval between nodes equal to the total evolutionary time divided by n, where n is the number of branching times. The relative propensities for amino acids to be in any location in the protein sequence as well as the local-structure-dependent mutation matrices used to simulate the site mutations were drawn from our earlier work (Koshi and Goldstein 1995). Given the amino acids that resulted in this procedure at the leaves of the evolutionary tree, we used equation 1 to ascertain the originally created ancestral protein.

One strength of this probabilistic analysis is our ability to represent degrees of certainty and include various possibilities along with the computed probabilities. First, however, we took the most likely ancestral amino acid as our prediction at that location and compared it with the amino acid that actually existed at that location in our model. Figure 2 shows the accuracy of our reconstruction, as a function of total evolutionary distance between root and leaves, for 4, 8, 16, and 32 present-day ho-

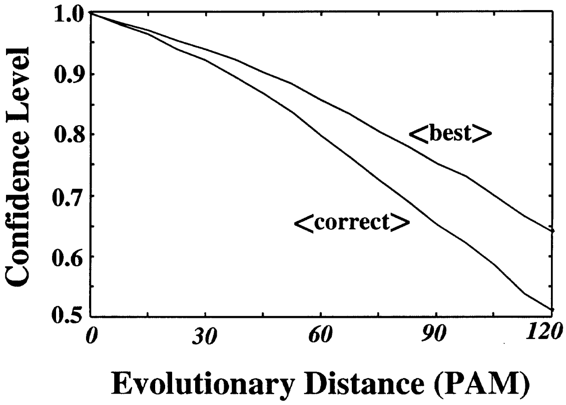

Fig. 3. For the 16 current homolog cases, the average confidence level of the predicted amino acid (⟨best⟩) and the average of the correct amino acid (⟨correct⟩) are plotted vs evolutionary distance (units in point-accepted mutations). The decline of the confidence level as the evolutionary distance increases mirrors the similar decline in the accuracy of the reconstruction, shown in Fig. 2.

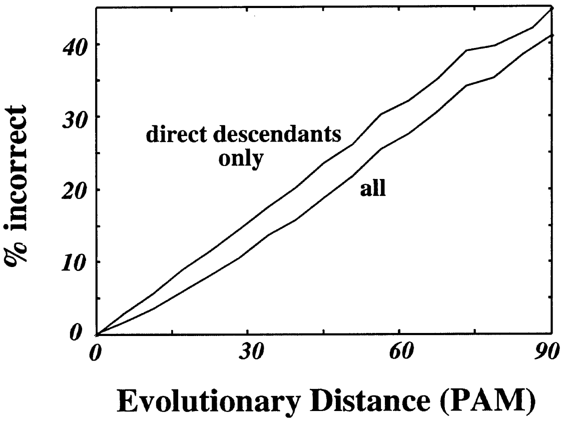

Fig. 4. Using 16 homologs, the sequence at one of the nodes directly below the root node was computed, and compared to the correct sequence at that point. Percentage incorrect vs evolutionary distance from that node to the bottom of the tree is plotted for two cases: using only direct descendants of that node to derive the sequence and using all homologs to derive the sequence.

mologs, assuming that the evolutionary tree and secondary structure at each point in the protein sequences are known exactly. As could be expected, the accuracy of the reconstruction drops off sharply with increasing evolutionary distance and increases when more homologs are available.

In our further testing, we focused on trees with 16 current homologs. Using 10 proteins each 500 residues long as a data set, we found the average probability assigned to the predicted ancestral residue as well as the average probability assigned to the correct ancestral residue. Figure 3 shows these results.

As discussed in the Introduction, this approach can be used to reconstruct any node, not just the root node of the tree. In particular, Fig. 4 shows the accuracy of the reconstruction of an ancestral node that is filial to the root of the phylogenetic tree both when information about all of the present-day homologous proteins is used and when only using information about direct descendents. As shown, useful information can be derived from a knowledge of other parts of the tree that are related to a node paternal to the node of interest.

The construction of the multiple sequence alignment and evolutionary tree structure must be computed prior to the use of this reconstruction methodology. The data in Fig. 2 assume that this has been done perfectly so that the true evolutionary tree and alignment are available. While there has been much progress in this area, there are still many unresolved issues. In order to evaluate how sensitive our technique is to the current limitations in this technology, we realigned the model sets of homologous proteins with ClustalV (Higgins et al. 1992), recomputed the evolutionary tree, and used these data in our reconstruction. The results are shown in Fig. 5, again for the example of 16 current homologous proteins. Also shown in Fig. 5 is the accuracy of the reconstruction using the ClustalV-generated tree when the protein structure is not known and a structure-independent mutation matrix is used.

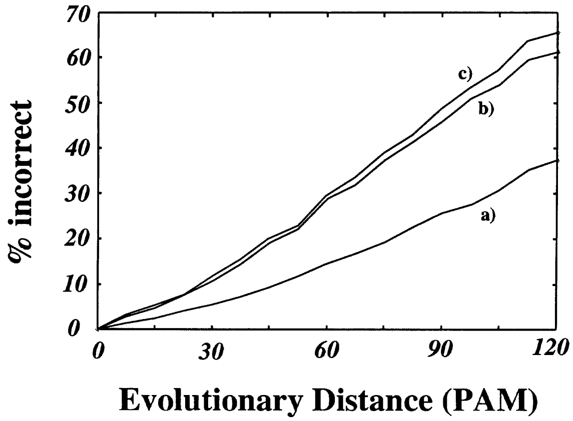

Fig. 5. In the 16-homolog case, percentage incorrect in the ancestral sequence vs evolutionary distance from the top to the bottom of the tree was plotted for three cases: *a*) using the known evolutionary tree and using-structure dependent matrices; *b*) using the tree generated by ClustalV from the alignments and using structure-dependent mutation matrices; and *c*) using the ClustalV tree and only a structure-independent matrix.

Finally, we applied these techniques to ribonuclease. With a test set of 38 proteins from the ribonuclease family, we formed an alignment and tree with ClustalV using an unrelated sequence to root the tree. The evolutionary tree which was derived is shown in Fig. 6. The average evolutionary distance between consecutive nodes in this tree corresponded to a sequence divergence of 7.8%, justifying our previous assumption of no double mutations between consecutive nodes. We then applied our methodology to reconstruct the ancestral sequence, using structure-dependent mutation matrices based on the local structure of bovine pancreatic ribonuclease (Brookhaven

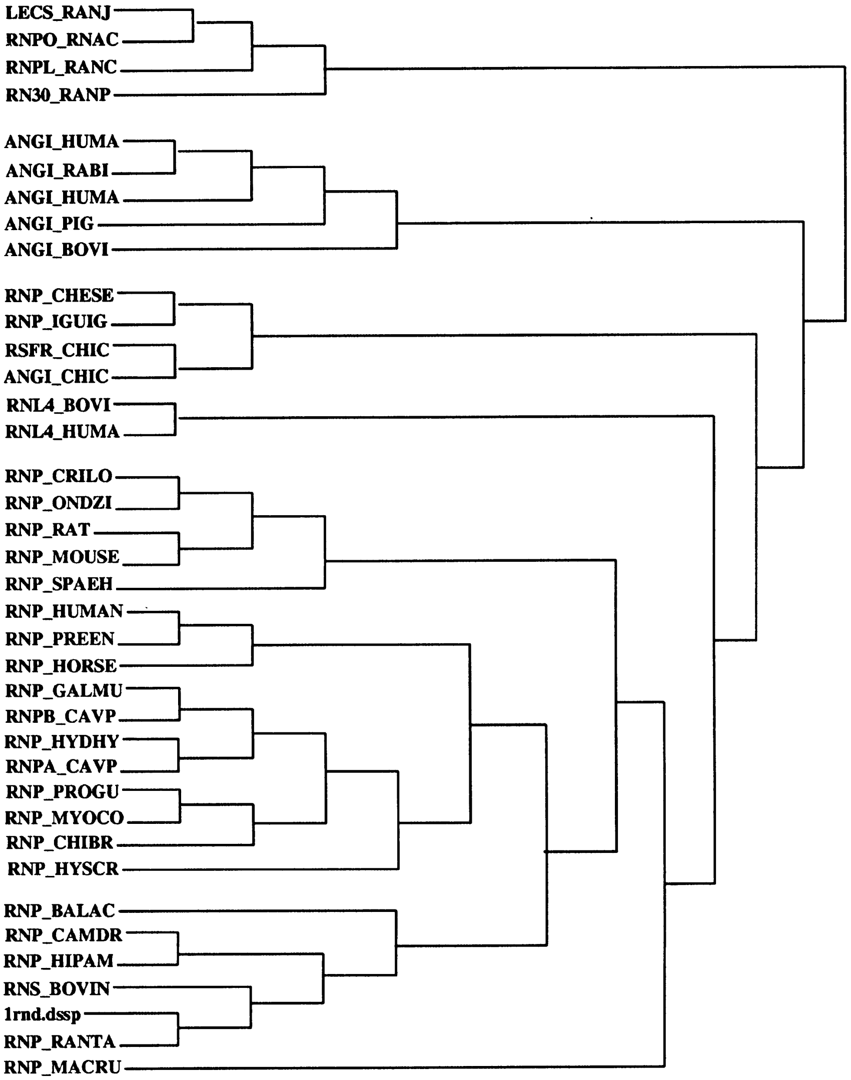

Fig. 6. Graphical representation of the tree generated by ClustalV used to reconstruct the ancestral ribonuclease sequence. (Evolutionary distances are not shown to scale.)

Protein Database designation 1RND). The results of this reconstruction, which took under a minute of CPU time on an SGI R4400, are shown graphically in Fig. 7 and numerically in Fig. 8.

Discussion

Ancestral reconstructions become more problematic as the evolutionary time increases and the number of homologs decreases. There are fundamental uncertainties that cannot be resolved without direct observation of ancestral protein or DNA sequences. Given these limitations, it becomes all the more important to be able to use all of the information available, such as knowledge of the structure of one of the members of the homologous set. And as the ability to furnish an exactly correct reconstruction decreases, it also becomes especially important to provide alternative possibilities with their respective probabilities. As is seen by a comparison of Figs. 2 and 3, as the accuracy of the reconstruction decreases, the probability reported by the analysis about the reconstruction decreases similarly. For 16 present-day homologs at an evolutionary distance of 120 PAM, the average confidence of the prediction, as represented by the probability assigned to the most-probable amino acid, is 64%. In fact, the reconstruction under these conditions has an accuracy rate of 63%. Certain locations in the sequence will be more constrained than others, and the reconstruction will be more accurate at these points—this increase in accuracy will be reflected by the calculated probabilities. The confidence level of correct predictions averaged 74%, while the confidence of predictions that were erroneous averaged only 48%. Generally, the correct ancestral residue had appreciable probability even in those cases of an incorrect prediction, when another residue had a still-higher probability. Even for the incorrect predictions, the confidence level of the correct answer averaged 13%.

This method is rather general, applicable for arbi-

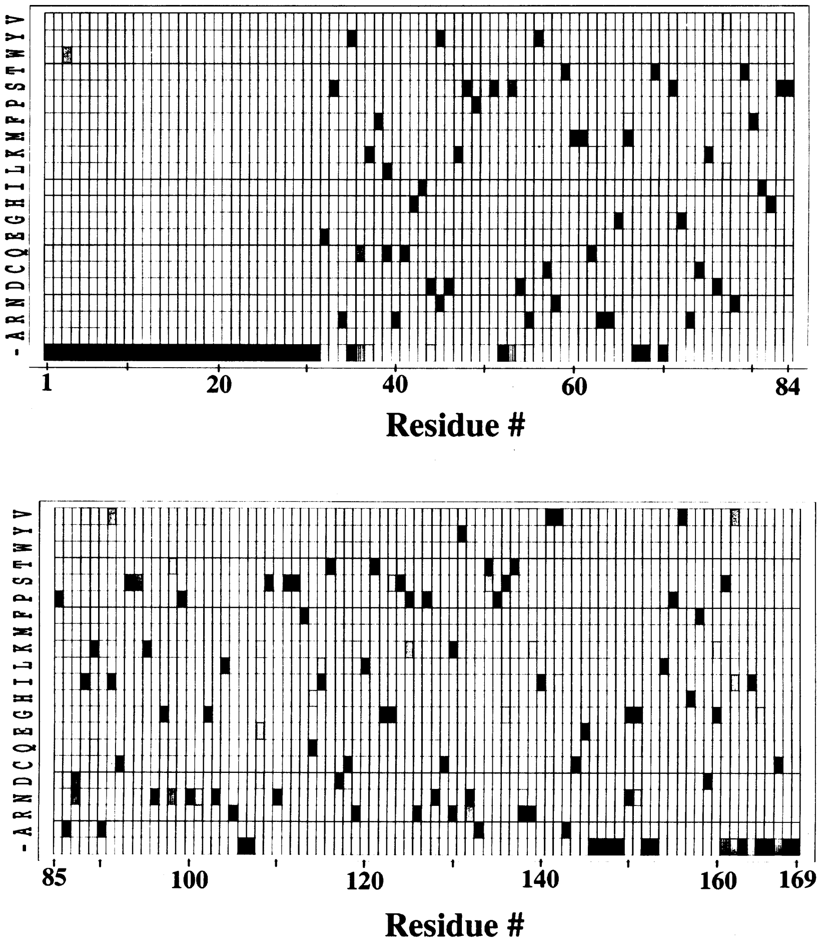

Fig. 7. Graphical representation of the probability of finding each amino acid in the ancestral protein to the ribonuclease superfamily. The shading of any square represents the probability of any amino acid having existed at that location in the ancestral protein sequence. *Black* represents a probability of one. Amino acids are abbreviated to their appropriate one letter codes, with gaps represented as –.

```
RN30_RANP   QDWLT--------------------------------FQKKHITNTRDVD------CDNIMSTNLF------HCKDKNTFIYSR
ANGI_BOVI   --------------------------AQD--DYRYIHFLTQH--YDAKPKGRNDEYCFNMMKNRRL--T-RP-CKDRNTFIHGN
RNP_CHESE   -------------------------------ETRYEKFLRQHVDYPKSSAPDSRTYCNQMMQRRGM--T-SPVCKFTNTFVHAS
RNP_CRILO   ---------------------VQPSLG---KESAAMKFERQHMDSTVATSS-SPTYCNQMMKRRNM--T-QGQCKPVNTFVHES

ancestral   -------------------------------ESR--KFLRQHIDYDKSPDS--DRYCNTMMQRRGM--T-SGRCKDLNTFIHSS
  -prob     776889999999999999999999999999978843695889755487628945499978978786997989699729999956
  -struc    ...............................  HHHHhHHhh          .  hhHHhhHH    .. .      Eeeee

RN30_RANP   PEPVKAICKGIIASKNV------LTTSEFYLSDC---NVTSRP-CKYKLKKSTNKFCVTCE----NQ--APVHFVG---VGSC--
ANGI_BOVI   KNDIKAICEDRNGQPYRGDLR--ISKSEFQITICKHKGGSSRPPCRYGATEDSRVIVVGCE----NG--LPVHFDES-FITRPH-
RNP_CHESE   AASITTVCGS-GGTPASGDLR--DSNASFALTTCRLQGGSQTPNCPYNADASTQRIRIACV----GG--LPVHYDKSI-------
RNP_CRILO   LADVHAVCSQENVKCKNGKSNCYKSHSALHITDCRLKGNAKYPNCDYQTSQHQKHIIVACE----GNPFVPVHFDATV-------

ancestral   PANIKAICSSKNGNPNNGNLR--ESNSSFQITDCRLTGGSPRPNCKYNATPSTRRIVVACE----GG--LPVHFDGS--I--C--
  -prob     9747796994798478387789939799966899879856599895959588977979899999955999999987439678599
  -struc    HHHhH     EEE        eeE    eeeeeeEE          eEeEEEEE eeeeeE....    eeEEeEEE .......
```

Fig. 8. Reconstructed ancestral ribonuclease. The first four sequences are examples chosen from the set used to create the ancestral sequence (see Fig. 6 for their location in the tree), and the single line of amino acids below (*ancestral*) is the reconstructed sequence of highest probability. The numbers below the ancestral sequence (*prob*) represent 10 times the probability of finding that amino acid in that position (truncated to the nearest multiple of 10). Local structure is represented by the bottom row of letters (*struc*)—*H* and *h* represent exposed and buried α-helix residues, respectively, while *E* and *e* represent exposed and buried β-sheet residues. *Dots* indicate positions where no knowledge of the local structure exists.

trarily complicated phylogenetic trees, and for any node in the tree. In particular, Fig. 4 demonstrates our ability to take advantage of our knowledge of sequences that are not direct descendents of the node in question. For the node being reconstructed, filial to the root node, only 8 of the 16 present-day sequences are direct descendents. Yet including information about the other eight sequences helps to boost the accuracy of the reconstruction.

In this methodology, we assume that we know the evolutionary tree *T*, and Fig. 5 demonstrates that this reconstruction is dependent upon an accurate phylogenetic tree. The phylogenetic analysis can be assisted by taking advantage of ancillary information that exists from other sets of homologous proteins, fossil records, and morphogenic comparisons. If the evolutionary tree is a matter of some uncertainty, this uncertainty can be naturally incorporated into the re-creation scheme. If $P(T_j)$ is the probability of evolutionary tree T_j, as computed for instance, using maximum-likelihood approaches, then equation 1 can be generalized to:

$$P(A_r|\{A_i\}', M) = \sum_j P(A_r|\{A_i\}', M, T_j)P(T_j) \qquad (7)$$

Uncertainty in the alignment of the homologous proteins can be handled in a similar way.

Figure 5 also shows that the accuracy of the reconstruction suffers when the local protein structure is unknown and structure-independent mutation matrices are used. Even when the structure of one of the members of the homologous set is known, uncertainties about the local structure will be caused by fluctuations of this structure during the evolutionary process (Rost et al. 1994; Rost and Sander 1994). In the complete absence of structural information, the local-structure-dependent mutation matrices can still be used to advantage. For instance, various secondary-structure prediction schemes can be used to estimate the probability of any given local structure. Less than perfect accuracy would still assist in the reconstruction. In addition, the observed evolutionary pattern itself provides information about the relative probability of various secondary structures. We can see this by considering $P(A_r,2_k^{\phi}\{A_i\}',T)$, the conditional probability that the root node has amino acid A_r *and* this location has secondary structure 2_k^{ϕ}, given residues $\{A_i\}'$ in current sequences and tree structure T.

$$P(A_r, 2_k^{\phi}|\{A_i\}', T) =$$

$$\frac{P(\{A_i\}'|A_r, 2_k^{\phi}, T)P(A_r|2_k^{\phi})P(2_k^{\phi})}{P(\{A_i\}'|T)} \qquad (8)$$

where 2_k^{ϕ} specifies the appropriate mutation matrix. The value of $P(A_r\{A_i\}',T)$ is then computed by summing $P(A_r,2_k^{\phi}\{A_i\}',T)$ over all possible secondary structures. Secondary structures with mutational patterns most consistent with the observed evolutionary process would be more heavily weighted in this sum. This approach can be used for any distribution of mutation matrices or a priori probabilities that depend upon unknown factors, such as in the modeling of heterogeneous mutation rates.

Figure 7 graphically demonstrates the ability of our method to provide probabilities of various possible reconstructions, options that can be possibly experimentally tested. The re-creation is in general rather confident (confidence of the predicted residue averages 83%) even given the rather long evolutionary time involved. The reconstruction is especially accurate for buried structural elements such as α-helices and β-sheets: such locations averaged 94% confidence. This reflects the lower mutation rates for the locations more critical for the protein structure and function.

Conclusion

The tracing of molecular phylogenies has become an increasingly useful approach in addressing major issues in evolution and biochemistry. With the advances in molecular biology, it has become possible to experimentally test possible scenarios by re-creating ancestral proteins and measuring their activity and stability. Given this new power, and the results of Benner and co-workers, which suggest that it is only at short time scales that nucleic acid substitution rates dominate mutation rates (Benner et al. 1994b), it becomes increasingly important to develop reconstruction schemes that work at the amino acid level and that can provide more than just the most probable sequence to test.

In this paper we have presented a method of ancestral reconstruction based upon an ML-type formalism and which is also based upon the longer-scale, amino acid level of molecular evolution. Our method is able to provide the probability of finding *any* amino acid at *any* point in the tree, and as the method is based upon amino acid substitutions, it is better able to incorporate the greater possibilities occurring at the residue level. Additionally, our method makes use of structure-dependent mutation matrices to include the information contained in secondary structure and surface accessibility.

Finally, we note that there has been interest in using information about correlated mutations to assist in the prediction of protein tertiary structure (Benner et al. 1994a; Gobel et al. 1994; Neher 1994; Shindyalov et al. 1994; Taylor and Hatrick 1994). Generally, these analyses have been limited to looking at correlations between the residues found in current-day proteins. With the reconstruction approach developed here, it may be possible to look at the correlations between the time-evolution of different residues at different locations.

Acknowledgments. We would like to thank Kurt Hillig for computational assistance. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, and NIH grants GM08270 and R29 LM05770.

References

Benner SA, Badcoe I, Cohen MA, Gerloff DL (1994a) Bona fide prediction of aspects of protein conformation. J Mol Biol 235:926–958

Benner SA, Cohen MA, Gerloff DL (1994b) Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7:1323–1332

Cooper, A, Mourer-Chauvire C, Chambers GK, von Haeseler A, Wilson AC, Paabo S (1992) Independent origins of New Zealand moas and kiwis. Proc Nat Acad Sci USA 89:8741–8744

Czelusniak J, Goodman M, Moncrief ND, Kehoe SM (1990) Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. Methods Enzymol 183:601–615

Dayhoff MO, Eck RV (1968) A model of evolutionary change in proteins. In: Dayhoff MO, Eck RV (eds) Atlas of protein sequence and structure, volume 3. National Biomedical Research Foundation Silver Spring, MD, pp 33–41

DeSalle R, Gatesy J, Wheeler W, Grimaldi D (1992) DNA sequences from a fossil: termite in oligo-miocene amber and their phylogenetic implications. Science 257:1933–1936

Felsenstein J (1973) Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Zool 20:406–416

Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18:309–317

Higgins DG, Bleasby AJ, Fuchs R (1992) Clustal V: improved software for multiple sequence alignment. CABIOS 8:189–191

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312:282–284

Holmquist R (1979) The method of parsimony: an experimental test and theoretical analysis of the adequacy of molecular restoration studies. J Mol Biol 135:939–958

Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices derived using Bayesian statistics and phylogenetic trees. Protein Eng 8:641–645

Libertini G, Donato AD (1994) Reconstruction of ancestral sequences by the inferential method, a tool for protein engineering studies. J Mol Evol 39:219–229

Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature 345:86–88

Moore GW, Barnabas J, Goodman M (1973) A method for constructing maximum parsimony ancestral amino acid sequences on a given network. J Theor Biol 38:459–485

Neher E (1994) How frequent are correlated changes in families of protein sequences. Proc Nat Acad Sci USA 91:98–102

Paabo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Nat Acad Sci USA 86:1939–1943

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226

Rost B, Sander C, Schneider R (1994) Redefining the goals of protein secondary structure prediction. J Mol Biol 235:13–26

Saitou N (1990) Maximum likelihood methods. Methods Enzymol 183: 584–598

Shih P, Malcolm BA, Rosenberg S, Kirsch JF, Wilson AC (1993) Reconstruction and testing ancestral proteins. Methods Enzymol 224:576–590

Shindyalov I, Kochanov N, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. Protein Eng 7(3):349–358

Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA (1990) The ribonuclease from an extinct ruminant. FEBS Lett 262: 104–106

Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. Protein Eng 7:341–348

Yang Z (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Systematic Biol 43:329–342