RC22279 (W0112-028) December 11, 2001

IBM Research Report

Error Analysis of Point Set Alignments for Use In Identification of Conserved Polypeptide Segments

Daniel E. Platt, Yuan Gao, Laxmi Parida, Isidore Rigoutsos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Identification of Conserved Polypeptide Segments with Spacial Clustering with Error Analysis of Point Set Alignments

Daniel E. Platt*, Yuan Gao,
Laxmi Parida, Isidore Rigoutsos
Bioinformatics and Pattern Discovery Group,
IBM Thomas J. Watson Research Center,
Yorktown Hgts, NY 10598,
http://www.research.ibm.com/bioinformatics

December 6, 2001

Abstract

A recognition of conserved protein structure frequently requires a three dimensional alignment of some of the common structural features of pairs of polypeptide segments represented as rigid point sets. The need for an estimation of a "mean" or representative structure composed from a collection of several possibly overlapping polypeptide segments also arises naturally in the context of such studies. There is a number of least-squares based solutions which have been exhaustively compared and described. However, techniques for computing

*contact: (914)945-1388, watplatt@us.ibm.com

a propagation of errors have not been developed. Such methods will provide ways of characterizing the variations tolerated by evolutionary conservation and diversity that may be observed in examples of conserved protein segments. Also, they will provide the means for testing whether new conservation candidates produce alignments consistent with the variability of previous observations.

This paper presents a computation for the propagation of errors for the χ^2 estimation of displacements and (orthogonally constrained) rotation matrices that comprise pose transformations. Further, a consistent formulation of "mean" or average point-sets constructed from an alignment of a group of rigid point sets is developed from the perspective of a minimization of a χ^2, together with an error analysis of the average point-sets along with the interactions with the translation and rotation errors of each of the rigid point sets to the average point set.

1 Introduction

There are a number of areas where the problem of characterizing the three-dimensional structural alignment of protein segments emerges naturally. The problem that motivated this paper is that of recognizing how the conserved peptide bases in divergent evolution are important to enzyme or protein function or conformation. Organisms do not survive mutations of bases that are important to enzymatic function. Only mutations of unimportant features which do not modify activity or which do not adversely inhibit conformational folding in the structure of the active sites remain viable. This implies that structures that are important to function will be conserved, suggesting that algorithms that seek to discover sequential patterns which appear more than once in a database can be used to identify candidates of evolutionary conservation.[1] The three dimensional structures matching those patterns may be clustered to identify putative examples of conservation which match the patterns, and recognize when such pattern matches arose simply by chance. Further, examples of conserved polypeptide sequences may be identified, and the variation tolerated by evolutionary processes may be characterized.

The specificity of function and structure promoted by divergent evolution stands somewhat opposed to the problem of protein structure prediction. The premise in structure prediction is that secondary and tertiary structure can be predicted entirly from the primary structure. To that extent, motifs that emerge as predictors of tertiary structure that cross functional or family

lines are the most useful. Originally, this study focused on seeking to identify patterns that could be used as predictors of protein structure.

On the other hand, the annotation problem seeks to identify those characteristics of the primary structure that are specific to functional or family lines. It is well known and understood that local sequence segments do not determine the immediate tertiary structure of the peptide sequences, and that there are many interactions with distant peptides that become important to the folding process which determine the tertiary structure contained within some local sequence segment. Since the effects of divergent evolutionary selection play such a dominant role in determing the tertiary conformation of a pattern segment, this study became much more focused on understanding the relationships between discovered patterns, conserved tertiary structures of matching sequences, and function.

This paper presents an analysis of those variations and tolerances associated with the alignments between protein segments required to achieve a clustering of 3D structures, and it presents a definition of an "average" strand produced by that clustering. First, a dictionary of patterns, called a "biodictionary," representing putative conserved peptide segments, was identified using the Teiresias[1] pattern discovery algorithm. Second, given some particular pattern, all the matches to that pattern provide atom-atom correspondences between the $C - C_\alpha - N$ on the backbone are extracted for input to a clustering algorithm. Since clustering by three-dimensional alignment requires some statement of tolerances, the interaction between those tolerances and the alignment process and discrimination of functionality should

be understood. Also, the specification of tolerances should reflect the uncertainties inherent in experimental techniques such as X-ray crystallography. The clustering is performed with the Jarvis-Patrick algorithm[2] with modifications outlined in appendix–C. The error analysis of the alignment problem is presented in appendix–A. Third, for a given cluster, a "mean" strand is defined in terms of a minimization of the RMSD of the multiple-alignment of all of the cluster members with the mean, together with the error propagation of the alignment parameters and the mean strand atom locations. This methodology is applied to a test case to explore the relationship between the cluster analysis tolerance parameters on the uncertainties in the alignment parameters and on the resulting specificity of functional categorization associated with the clusters.

Teiresias is a pattern discovery method[1] that identifies all repeated patterns present in a given set of protein sequences without requiring alignment or enumeration of the actual matches. The discovered patterns match on over 98% of the peptide positions in the GenPept database, indicating that these matches may form a comprehensive set of descriptors indexing the peptide sequence space. As pointed out before, any pattern conserved by divergent evolutionary processes present in the database should emerge from such a discovery algorithm. However, such discovered patterns may detect patterns matching multiple evolutionary paths (convergent evolution in sequence patterns if not geometry and function), or matches by chance. Relatively dense patterns are mandated since the size of enumerable space indexed by the discovered patterns increases rapidly with decreaseing density.

The distance metric used in the modified Jarvis-Patrick algorithm is a straightforward RMSD χ^2 variate, which with an assumed atomic uncertainty radius, reduces to a simple RMSD between the best alignment of the $C - C_\alpha - N$ atoms in each peptide match along the protein backbone associated with each other by the matching patterns. The fact that the discovered patterns are used simplifies the problem definition because the requisite atom-atom associations matching the discovered patterns are guaranteed to be unambiguous. The modification to the Jarvis-Patrick algorithm is that molecular fragments with best-aligned RMSDs larger than a certain cutoff are not considered to be neighbors. One of the specific patterns matching a segment in a sample query was selected.

There seems to be low homological matching to the selected query, 2hsd. This makes it an interesting test case since its low homology should cause it to show numerous pattern matches which do not reflect conservation. In this way, the selected seqlets should become something of a random sample of the biodictionary, rather than one that was dominated by a particular type of enzyme or function class. An exploration of the parameter space is performed on PDB correspondents to that specific pattern, with particular consideration to the known query structure. For this case, the effects of adjustments of clustering tolerances on the alignment uncertainties and upon the discrimination of functionality are explored.

Although the approach of this paper shares the intellectual underpinnings and motivations of other studies of 3 dimensional similarity in proteins, this study became concentrated on the specific problem of trying to identify which

candidate matches to discovered patterns are actually examples of conservation. Initially, this study emerged from consideration of the problem of identifying similarities between portions of known databases that matched patterns derived from a pattern discovery algorithm as possible candidates for structure prediction. This study focused on the use of clustering the 3D segments matching the patterns identified by the query, in order to identify which database matches represented candidates for conservation, and which were likely spurious matches. Therefore, this study most strongly resembles the track of study pursued by Jonassen et. al.[3, 4] While Jonnassen et. al. have considered pattern matches, which most closely resembles Teiresias, they have also pursued probabilistic matching schemes in their PROSITE study, with extensions to three-dimensional data. Their approach to three-dimensional characterization has involved the selection of "probe" candidates to derive scores screened by RMSD thresholds. That approach is distinct from the approach taken in this study.

This approach is also distinct from others which seek to identify candidates of conservation directly from geometry. For example, the computational basis for three dimensional similarity comparisons have included distance maps,[5] use of distance maps to classify domains recognized by other algorithms, as in FSSP,[6] classifications by direct human intervention, recognition being by eye including SCOP,[7] hybrids between automated and human constructed databases, including CATH.[8] All of these approaches consider similar questions such as what constitutes a "fold," how many folds are there,[9, 10, 11] as well as some characterization of their diversity.[12]

Some of these use detailed alignment between point-site associations, but many of them depend on more general alignments between secondary structure elements. While all of these seek to recognize similarities that may or may not reflect common evolutionary roots in the guise of super-families for the purposes of protein structure prediction, one study has shown that IF3, a protein involved in the initiation stage of protein synthesis, shows very similar structural features in *E. coli* and *B. stearothermophilus*, as well as between those features and function, but which show different stabilization mechanisms and linker regions that determine the folding process.[13] This highlights some of the difficulties of finding meaningfully conserved patterns.

More specifically, one study[14] considers the problem of identifying all possible partial alignments of protein elements with other proteins. That approach is distinct from the one presented in this discussion because that study sought possible correspondences between elements that may be separated by much larger distances than those that would be identified by a pattern discovery algorithm, while this study focuses on the recognition of candidates for conserved structure in short segments that have been identified by a pattern discovery algorithm. Nevertheless, the statistical characterization of alignment presented in this work depends only upon the provision of point-point associations. Since both the associated segments presented by pattern discovery and the partial alignment algorithms provide point-point associations, the alignment description is equally valid in both problem domains.

The general problem of identifying geometric conservation from among

discovered patterns, as outlined in this study, has some distinct characteristics compared to prior studies in that the correspondences are based on discovered pattern segments which are generally smaller than folds, and the point-point associations are derived specifically from the pattern correspondences or matches with the peptide segments. On one hand, the patterns are generally too small to relate secondary structure features to each other, such as groups of α helices that occasionally form pockets for active sites. The relationships between α helices can discriminate function very effectively. Typically, only parts of α helices are usually captured by the pattern segments, and the ubiquity of helices eliminates discrimination of function. On the other hand, only four or five bases representing a turn extending from an α helix are sufficient to be fairly discriminating of function. The methods described in this presentation open a window on functional discrimination at a distance scale that, to the best of our knowledge, is not developed in other studies.

2 Error Propagation of Alignment from χ^2 Considerations

A complete description of a "best alignment" should include an error analysis that specifies a range of tolerance implied by the variation in the data. Least-squares approaches to parameter estimation are intuitively appealing, and they provide a measure of the quality of the alignment according to the RMSD. The problem of pose estimation is essentially identical to a regression problem with quadratic constraints.

This section presents the reasoning employed in using χ^2 for the purposes of error propagation as applied in parameter estimation and regression problems.

Two types of alignment problems are considered here. First, pair-wise alignment appropriate for identifying the relationship between a querry and a retrieved target is considered in appendix–A. This step is necessary in the "collection recognition" or clustering phase of the problem where best estimates of RMSD's must be computed. Second, alignment between multiple targets with an estimated collection mean together with its error analysis is presented in appendix–B. This analysis provides a way of determining what the best estimates of an average or mean backbone consistent with the cluster members would be along with the error estimates for the mean. The detailed derivations in appendices A and B follow arguments essentially identical to that outlined below.

2.1 Error Propagation of Alignment

Consider N variables z_i which are normally distributed with mean 0 and standard deviation 1. Then a variable $u = \sum_{i=1}^{N} z_i^2$ is χ^2 distributed with N degrees of freedom.[15] If a collection of N variables x_i are distributed with means μ_i and standard deviations σ_i, then the variable $z_i = (x_i - \mu_i)/\sigma_i$ is normally distributed with mean 0 and standard deviation 1. This means χ^2 variables may be constructed that measure the consistency of a set of data x_i with their putative means μ_i given expected dispersions σ_i.

Traditionally, the most familiar application of the χ^2 statistic has been in the analysis of contingency tables. This application shows a χ^2 statistic $\mathcal{E}^2 = \sum_j (x_j - Np_j)^2/(Np_j)$ with $N - 1$ degrees of freedom, which does not overtly resemble the definition provided above. However, the definition of a χ^2 distributed variable is $\chi^2_N = \sum_{i=1}^{N} z_i^2$ where the z_i are independent normally distributed variables. It can be shown that for a multinomial distribution $f(\{x_j\}) = N!/(\prod_j x_j!) \prod_j p_j^{x_j}$, that $\mathcal{E}^2 = \sum_j (x_j - Np_j)^2/(Np_j)$ where $\sum_j x_j = N$ is approximately χ^2_{N-1} distributed, and it is for this reason that the χ^2 test of interactions between contingencies is applied.

The application of χ^2 ideas to error propagation may be shown more simply by considering the problem of estimating μ and σ where $(x_j - \mu)/\sigma$ are N independent and normally distributed variables. Then

$$\mathcal{E}^2 = \sum_j \frac{(x_j - \mu)^2}{\sigma^2}$$

is χ^2_N distributed.

It is then possible to expand $\mathcal{E}^2$ about μ to yield

$$\mathcal{E}^2 = \frac{N}{\sigma^2}\left\{\frac{1}{N}\sum_j x_j^2 - \left(\frac{1}{N}\sum_j x_j\right)^2\right\} + \frac{N}{\sigma^2}\left(\mu - \frac{1}{N}\sum_j x_j\right)^2.$$

The linear combination of x_j's will also be normally distributed. The fact that it is one of a quadratic contribution to a χ^2 distributed variable implies that it is also χ^2 with one degree of freedom. Thus

$$E\left(\frac{1}{N}\sum_j x_j\right) = \mu,$$

and

$$E\left\{\left(\frac{1}{N}\sum_j x_j - \mu\right)^2\right\} = \sigma^2/N,$$

accounting for the $1/N$ dependence in the variance of the sample mean.

Also, it is possible to partition $\mathcal{E}^2$ into two parts

$$\mathcal{E}^2 = \mathcal{E}^2_{\text{res}} + \mathcal{E}^2_{\text{param}},$$

where

$$\begin{aligned}
\mathcal{E}^2_{\text{res}} &= \frac{N}{\sigma^2}\left\{\frac{1}{N}\sum_j x_j^2 - \left(\frac{1}{N}\sum_j x_j\right)^2\right\}, \\
\mathcal{E}^2_{\text{param}} &= \frac{N}{\sigma^2}\left(\mu - \frac{1}{N}\sum_j x_j\right)^2,
\end{aligned}$$

with $\mathcal{E}^2_{\text{param}}$ representing all the information concerning the quality of the prediction of the parameter μ, and $\mathcal{E}^2_{\text{res}}$ representing the residual error.

Since $\mathcal{E}^2_{\text{param}}$ is χ^2 with one degree of freedom, and $\mathcal{E}^2$ is χ^2 with N degrees of freedom, it follows that $\mathcal{E}^2_{\text{res}}$ is χ^2 with $N-1$ degrees of freedom.

Since $E(\chi^2_{N-1}) = N-1$, it follows that

$$E(\mathcal{E}^2_{\text{res}}) = N-1 = \frac{N}{\sigma^2} E\left(\left\{\frac{1}{N}\sum_j x_j^2 - \left(\frac{1}{N}\sum_j x_j\right)^2\right\}\right),$$

or

$$\sigma^2 = \frac{N}{N-1} E\left(\left\{\frac{1}{N}\sum_j x_j^2 - \left(\frac{1}{N}\sum_j x_j\right)^2\right\}\right),$$

accounting for the $N-1$ in the unbiased estimator of σ^2. If σ was already known, then it would be possible to compute a probability $P(\mathcal{E}^2_{\text{res}} > \chi^2_{N-1})$ to compare with a confidence threshold.

2.2 Error Propagation of Alignment to an Average Backbone

The model for simple alignment is simply that of

$$y_{ij} = \sum_k a_{jk} x_{ik} + d_j + e_{ij}, \tag{1}$$

where y_{ij} is the j'th component of the target coordinates for the i'th point to be aligned with x_{ik} using a rotation specified by the matrix a_{jk} and the displacement d_j. The quality of the alignment is assumed to be subject to some variations which will be accounted for by a statistical error e_{ij}. It is assumed that the e_{ij} are independent and normally distributed with mean 0 and variance $E(e^2_{ij}) = \Delta y_i^2$. Then the goodness of fit

$$\mathcal{E}^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j e^2_{ij} = \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k a_{jk} x_{ik} + d_j - y_{ij}\right)^2$$

is χ^2 distributed. The analysis is handled much the same as above, except the variation will be about estimates A_{jk} and D_j of a_{jk} and d_j, so that

$$d_j \;=\; D_j + \delta d_j,$$

$$a_{jk} = A_{jk} + \delta a_{jk},$$

and where the a_{jk} and A_{jk} are rotation matrices subject to orthogonality constraints

$$\sum_j a_{jk} a_{jk'} = \delta_{kk'}.$$

The problem of defining a mean in a meaningful way may perhaps be best addressed by noting that in the above analysis, $\mathcal{E}^2_{\text{param}}$ minimizes where $\mu = 1/N \sum_j x_j$. In this sense, the above example is a one-dimensional alignment with a mean, and it provides an error analysis of that alignment.[1] However, y_{ij} becomes the j'th coordinate of the i'th point in the mean, and it is now a parameter whose quality is subject to a χ^2 expansion. There are also now multiple segments aligned with the mean, according to the relationship

$$y_{ij} = \sum_k a_{ljk} x_{lik} + d_{lj} + e_{lij}$$

where y_{ij} is the average of the l points i subject to alignments described by a_{ljk}, d_{lj}. In this case, the χ^2 distributed error

$$\mathcal{E}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j e_{lij}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j \left(\sum_k a_{ljk} x_{lik} + d_{lj} - y_{ij} \right)^2$$

is expanded about all the parameters

$$\begin{aligned} d_{lj} &= D_{lj} + \delta d_{lj}, \\ a_{ljk} &= A_{ljk} + \delta a_{ljk}, \\ y_{jk} &= Y_{jk} + \delta y_{jk}, \end{aligned}$$

[1] A more complete analysis that includes rotations and displacements of three dimensional point sets is outlined in appendix–B. The basic ideas are essentially simple extensions of the analysis in appendix–A.

where the a_{ljk} are subject to the same orthogonality constraints

$$\sum_j a_{ljk} a_{ljk'} = \delta_{kk'}.$$

Expansions of $\mathcal{E}^2$ about A_{ljk}, D_{lj}, and Y_{jk} yields a statistical characterization of how well a cluster may by parameterized by an average point set, as well as how those parameterizations depend on the tolerances in alignment implied by the looseness of the best alignments.

It should be noted that the estimates of the error bars have the same character as the $1/N$ dependence of the sample mean variation described above rather than that of a sample-space variation. It is therefore possible to compute uncertainties implied by rigid alignments for points that were not included in alignments computed from some sub-sample of points. The quality of an alignment for excluded points as a test of consistency, perhaps to test whether other important conserved areas are consistent with the variations of the sub-sample, should be done using a computation of $\mathcal{E}^2$ using the estimates of ΔY^2 obtained from $\mathcal{E}^2_{\rm res}$. Further, it is possible to test whether some new peptide segment is consistent with one or another cluster of peptide segments by computing a $P(\mathcal{E}^2_{\rm res} > \chi^2_{N-1})$ for the unclustered segment.

2.3 Clustering

This section considers how the cluster parameters interact with alignment error propagation. The clustering algorithm parameters specify several classes of tolerances limiting cluster membership. These parameters find expression in the uncertainties in alignment between putative members and in the un-

certainty in a representative or mean segment. Further, these tolerances also affect specificity in recognizing and identifying possible function. A more detailed study of the relationship between sequence patterns conserving three-dimensional structure and function will be reserved for another paper. In the discussion that follows, this paper will consider some of the issues involving this specificity.

A collection of amino acid sequence patterns was produced using the Teiresias pattern discovery algorithm.[1] The product of the algorithm satisfies the following specifications: the algorithm identifies patterns, called "seqlets," characterized by a minimum number L of "literals" (i.e. non-wild-card characters) and some number of wild-card characters that are allowed to match any peptide; any consecutive L literals in the reported seqlets span at most W positions (typically, the length of reported seqlets is much longer than W); all seqlets must begin and end with a literal. For a choice of the values L and W, the reported seqlets are guaranteed to satisfy the following conditions: a) they span at least L positions; b) of any consecutive L literals (note: consecutive does *not* mean contiguous) in a reported seqlet, the first and last literals are not more than W positions apart; and c) if a reported seqlet appears x times in the input, it is guaranteed that a more specific seqlet cannot be constructed through appending or prepending another string, or through dereferencing a wild-card character without either decreasing the number of its occurrences or violating the L and W constraints. This algorithm was applied to the sequences reported in the PDB database. The resulting list, or dictionary, of seqlets is called a "biodictionary."

For each instance of a biodictionary seqlet entry with a correspondence in the query, the α-chain backbone coordinates of corresponding peptide sequence fragments found in the PDB database were extracted and clustered according to the algorithm described in appendix–C. The clustering algorithm considers two items (as graph nodes) to be connected (as in graph edges) to each other if they share $K_{\max}$ of each others' nearest K neighbors using RMSD as a metric, if each one is in the first K of the other's nearest neighbors, and if the are within $R_{\max}$ RMSD of each other. Then a cluster is constructed of all the members which are connected by some path through the edges.

3 Application to Examples

The Teiresias pattern discovery algorithm was applied to the GenPept database of November 15, 1999, using parameters $L = 6$, $W = 15$. These discovered peptide sequence patterns are candidates for structural conservation because they satisfy the necessary condition that they repeat in the database of processed protein sequences, and while there is some chance any one match to a pattern might have been spurious, the probability that at least some of the matches are significant is high. This list of discovered patterns become a dictionary of candidates, also referred to as the "Biodictionary,"[1] for conserved markers of function and structure. The Patterns were matched to a querry subject to chemical equivalence substitutions listed in table 1.

Table 1: Chemical Equivalence Substitutions

$A \to A$, $C \to C$, $D \to D$, $E \to D$, $F \to F$, $G \to A$, $H \to H$, $I \to I$, $K \to K$, $L \to I$, $M \to I$, $N \to N$, $P \to P$, $Q \to N$, $R \to K$, $S \to S$, $T \to S$, $V \to I$, $W \to W$, $Y \to F$

There were 6458 instances of correspondences to patterns in the dictionary with fragments of the querry, the 2hsd peptide sequence obtained from the PDB, Release 90 (see table 2).

Clustering was performed with two sets of parameters while only varying the distance cutoff parameter $R_{\max}$ for tolerance comparison purposes. The first parameter set was $K = 8$ elements considered in each nearest neighbor list, at least $K_{\max} = 2$ common or shared nearest neighbor entries, a maximum RMSD of $R_{\max} = 1.0$ $\mathring{\text{A}}$, and the second parameter set was

Table 2: 2HSD – $3\alpha, 20\beta$-Hydroxysteroid Dehydrogenase (Holo Form)

```
NDLSGKTVII TGGARGLGAE AARQAVAAGA RVVLADVLDE EGAATARELG
DAARYQHLDV TIEEDWQRVV AYAREEFGSV DGLVNNAGIS TGMFLETESV
ERFRKVVEIN LTGVFIGMKT VIPAMKDAGG GSIVNISSAA GLMGLALTSS
YGASKWGVRG LSKLAAVELG TDRIRVNSVH PGMTYTPMTA ETGIRQGEGN
YPNTPMGRVG EPGEIAGAVV KLLSDTSSYV TGAELAVDGG WTTGPTVKYV
MGQ
```

$K = 8, K_{\max} = 2, R_{\max} = 2.5$ Å. The 3D structure segments matching the peptide dictionary pattern $A.....A.A..VLA$ was selected as a representative example. The 3D segments matching the pattern are shown in Figure 1.

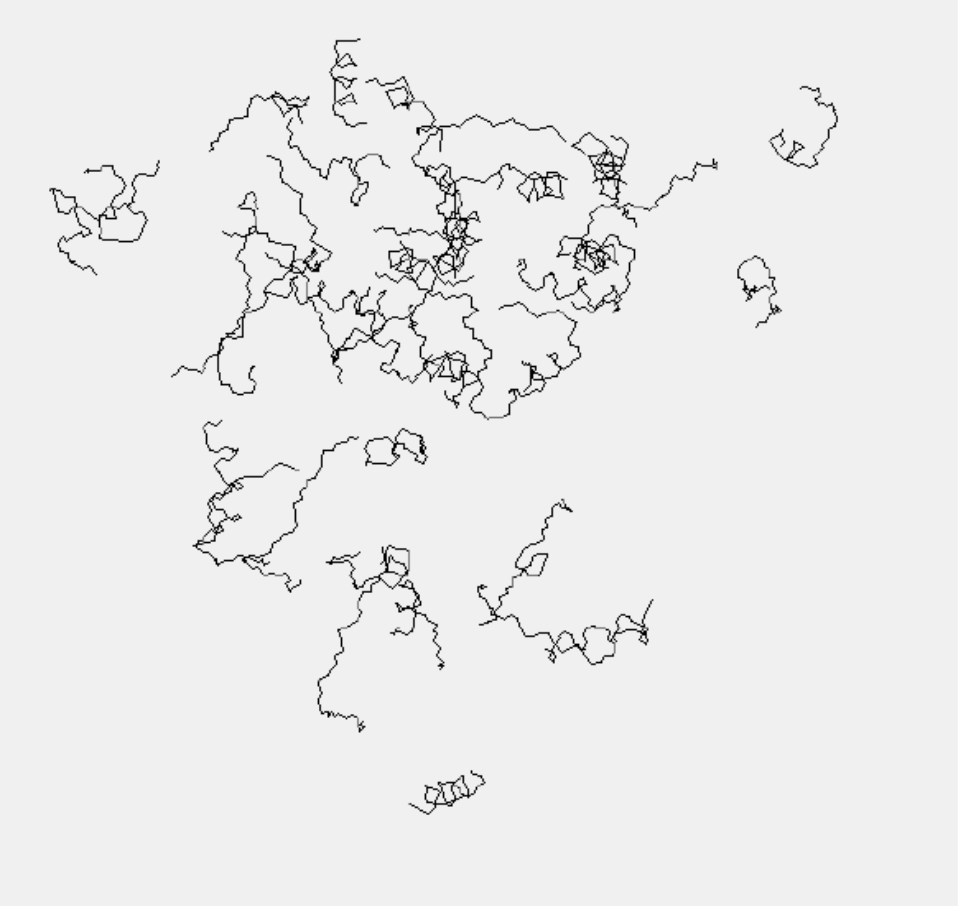

Figure 1: Matches to pattern, not clustered or aligned

Table 3: A.....A.A..VLA Cluster Members

RCSB Entry	Description
1aqn	Serine Protease (Subtilisin), Bacillus Amyloliquefaciens
2hsd	Oxidoreductase (Hydroxysteroid Dehydrogenase), Streptomyces Exfoliatus
2sic	Proteinase (Subtilisin), Bacillus Amyloliquefaciens
1arg	Transferase (Aspartate Aminotransferase), Escherichia Coli
1cov	Coxsackievirus Coat Protein, Coxsackievirus
1dpt	Cytokine (D-Dopachrome Tautomerase), Homo Sapiens
1mdt	Toxin (Monomeric Diphtheria Toxin), Corynebacterium Diphtheriae
1prc	Photosynthetic Reaction Center, Blastochloris Viridis
2trs	Lyase (Tryptophan Synthase), Salmonella Typhimurium
1b3n	Condensing Enzymes (Ketoacyl Acyl Carrier Protein Synthase), Escherichia Coli
1efp	Electron Transport (Electron Transfer Flavoprotein), Paracoccus Denitrificans
1cdo	Oxidoreductase (Ch-Oh(D)-Nad(A))(Alcohol Dehydrogenase), Gadus Callarias
1deh	Oxidoreductase (Human β 1 Alcohol Dehydrogenase), Homo Sapiens

Of the 44 3D segments extracted from the PDB that matched this pattern, only four clusters of three, six, two, and two entries respectively matched this sequence out of the first more restricted tolerance set, listed in table 3. Figure 2 shows the unaligned members from the cluster of six segments.

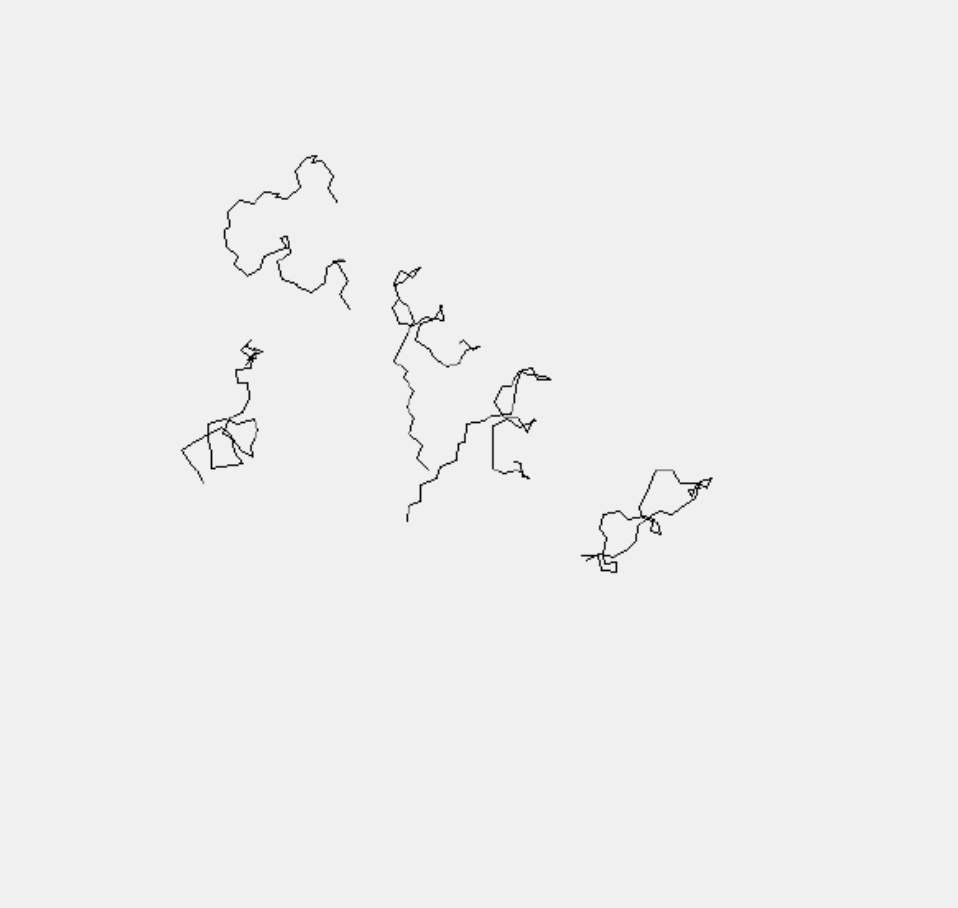

Figure 2: Clustered matches to pattern, not aligned

The following issues concerning function specificity have been observed, but a more detailed and systematic study is reserved for later publication. Frequently, more than one cluster may be extracted from three dimensional peptide sequence conformations matching a sequence pattern obtained from the PDB database. Limited by the window size of the discovered patterns, helices show very little specificity of function in isolation of any other information. The second cluster listed in table 3 is typical of this situation. The presence of almost any other kind of turn is frequently very specific in associating enzyme function with any particular cluster whose peptide sequences match a pattern.

The selected pattern shows two exceptions. First, 2hsd, a hydroxysteroid dehydrogenase, is not in the same class as 1aqn and 2sic, which *are* both subtilisins. Further, 1b3n, a ketoacyl acyl protein synthase, is not in the same functional group as 1efp, an electron transport flavoprotein. Functional similarities may be more effectively established if such proteins appear in the same clusters for multiple patterns.

The construction of multiple clusters from any particular pattern proves that unique identifications of the query enzyme type are not always guaranteed for any specific pattern, as is seen by multiple 3D clusters with correspondence to $A.....A.A..VLA$. This is both expected and desireable.[2] Generally, each of the clusters represents a candidate for conserved structure. Finally, while the sequences matching the pattern is expected to show clusters,

[2]Naturally, some patterns do have such properties and can be used as predicates of family membership (see also [1] for detailed discussion).

some matched sequences do not cluster with any of the other matched sequences. For example, another pattern from the dictionary, $A.....A.AR..LA$, shows correspondence in the query, 2hsd, but the corresponding 2hsd backbone segment does not cluster with any other 3D segments that match the pattern.

An average backbone was computed for each cluster according to the algorithm described in appendix–B, together with its coordinate errors, as listed in table 4. Finally, the pair-wise alignment of each peptide with the average backbone was computed according to the method outlined in appendix–A, and listed in table 5. Note that while the methods described in appendix–B also compute error bars on the transformation parameters $\mathbf{A}$ and $\vec{d}$ *as well as* their covariances with the uncertainties of the means. The entries listed in table 5 do not account for those dependences. Rather they are computed as simple alignments with a given point set as outlined in appendix–A. Figure 3 shows the best alignments of the cluster members with the mean, together with the uncertainties in the mean as enumerated in table 5. The sphere radii are set equal to the computed uncertainties. Those uncertainties, as tabulated in table 5, are essentially computed from $\sqrt{\sum_j E(\delta d_{ij}^2)}$, based on the variations within each cluster about the mean segment.

Table 4: A.....A.A..VLA Average Coordinates

x	y	z	ΔY
20.0339	32.0787	16.641	0.28636
19.0833	33.1219	16.1161	0.29115
17.8288	32.9911	16.5154	0.285795
16.8124	33.9006	16.0204	0.278043
17.0794	35.3025	16.5198	0.290087
17.4196	35.4171	17.7948	0.296162
17.7084	36.735	18.3966	0.300667
18.9219	37.3472	17.7473	0.300512
19.9503	36.5196	17.5362	0.294308
21.1943	36.9849	16.9341	0.283983
20.9417	37.4724	15.5153	0.293452
20.1213	36.7344	14.7921	0.299516
19.7688	37.0976	13.4133	0.299543
19.0071	38.4008	13.3572	0.299861
18.0859	38.5619	14.2976	0.298863
17.2672	39.7937	14.3317	0.295138
18.1784	40.9956	14.5915	0.302841
19.0896	40.843	15.5378	0.303733
19.9892	41.9656	15.8942	0.296247
20.8958	42.2509	14.6984	0.290792
21.2848	41.1903	13.9968	0.291463
22.1295	41.351	12.8041	0.285237
21.3663	42.0193	11.6884	0.293609
20.0633	41.9952	11.7629	0.298692
19.1799	42.584	10.7547	0.292949
18.7676	44.0096	11.1347	0.291903
19.2298	44.4989	12.2722	0.296579
18.9085	45.8607	12.6795	0.294348
17.809	46.0171	13.7083	0.294761
17.2525	44.9441	14.2367	0.295955
16.1706	45.0551	15.2212	0.292183
16.7477	45.2248	16.6215	0.299102
16.036	45.9655	17.4507	0.296423
16.3623	46.0937	18.8747	0.296634
15.5235	45.003	19.5549	0.297728
16.1977	44.0936	20.2198	0.299054
15.5258	42.9763	20.8929	0.299305
15.5686	43.1233	22.4068	0.295472
14.4543	42.9187	23.0165	0.29378
14.3184	43.0072	24.4467	0.289459
13.8153	41.6931	24.9388	0.287924
14.4429	41.2021	25.9344	0.283857
14.0023	39.9889	26.5071	0.281927

Table 5: A.....A.A..VLA Transformation Parameters

Parameter	Values		
	1aqn		
A	$1.00000 \pm 2.03009 \times 10^{-2}$	$-4.63312 \times 10^{-8} \pm 7.15983 \times 10^{-3}$	$-3.92743 \times 10^{-8} \pm 1.07539 \times 10^{-2}$
	$4.63312 \times 10^{-8} \pm 2.03009 \times 10^{-2}$	$1.00000 \pm 7.15983 \times 10^{-03}$	$6.58988 \times 10^{-8} \pm 1.07539 \times 10^{-2}$
	$3.92743 \times 10^{-8} \pm 2.03009 \times 10^{-2}$	$-6.58988 \times 10^{-8} \pm 7.15983 \times 10^{-3}$	$1.00000 \pm 1.07539 \times 10^{-2}$
$\vec{d}$	$2.53612 \times 10^{-6} \pm 3.00617 \times 10^{-2}$	$-1.93455 \times 10^{-6} \pm 3.00617 \times 10^{-2}$	$1.96735 \times 10^{-6} \pm 3.00617 \times 10^{-2}$
	2hsd		
A	$-8.24407 \times 10^{-1} \pm 2.59421 \times 10^{-2}$	$-4.10791 \times 10^{-1} \pm 2.61336 \times 10^{-2}$	$-3.89363 \times 10^{-1} \pm 3.44454 \times 10^{-2}$
	$-1.29506 \times 10^{-2} \pm 2.59421 \times 10^{-2}$	$7.01434 \times 10^{-1} \pm 2.61336 \times 10^{-2}$	$-7.12616 \times 10^{-1} \pm 3.44454 \times 10^{-2}$
	$5.65849 \times 10^{-1} \pm 2.59421 \times 10^{-2}$	$-5.82444 \times 10^{-1} \pm 2.61336 \times 10^{-2}$	$-5.83588 \times 10^{-1} \pm 3.44454 \times 10^{-2}$
$\vec{d}$	$8.14819 \times 10^{1} + / - 8.00576 \times 10^{-2}$	$7.24669 \times 10^{1} + / - 8.00576 \times 10^{-2}$	$-4.61818 + / - 8.00576 \times 10^{-2}$
	2sic		
A	$7.90993 \times 10^{-1} \pm 1.28578 \times 10^{-2}$	$-1.75299 \times 10^{-1} \pm 9.11184 \times 10^{-3}$	$5.86174 \times 10^{-1} \pm 2.29967 \times 10^{-2}$
	$-1.44383 \times 10^{-1} \pm 1.28578 \times 10^{-2}$	$8.77532 \times 10^{-1} \pm 9.11184 \times 10^{-3}$	$4.57265 \times 10^{-1} \pm 2.29967 \times 10^{-2}$
	$-5.94544 \times 10^{-1} \pm 1.28578 \times 10^{-2}$	$-4.46327 \times 10^{-1} \pm 9.11184 \times 10^{-3}$	$6.68812 \times 10^{-1} \pm 2.29967 \times 10^{-2}$
$\vec{d}$	$-1.35007 \times 10^{1} \pm 3.34316 \times 10^{-2}$	$5.96893 \pm 3.34316 \times 10^{-2}$	$3.96254 \times 10^{1} \pm 3.34316 \times 10^{-2}$

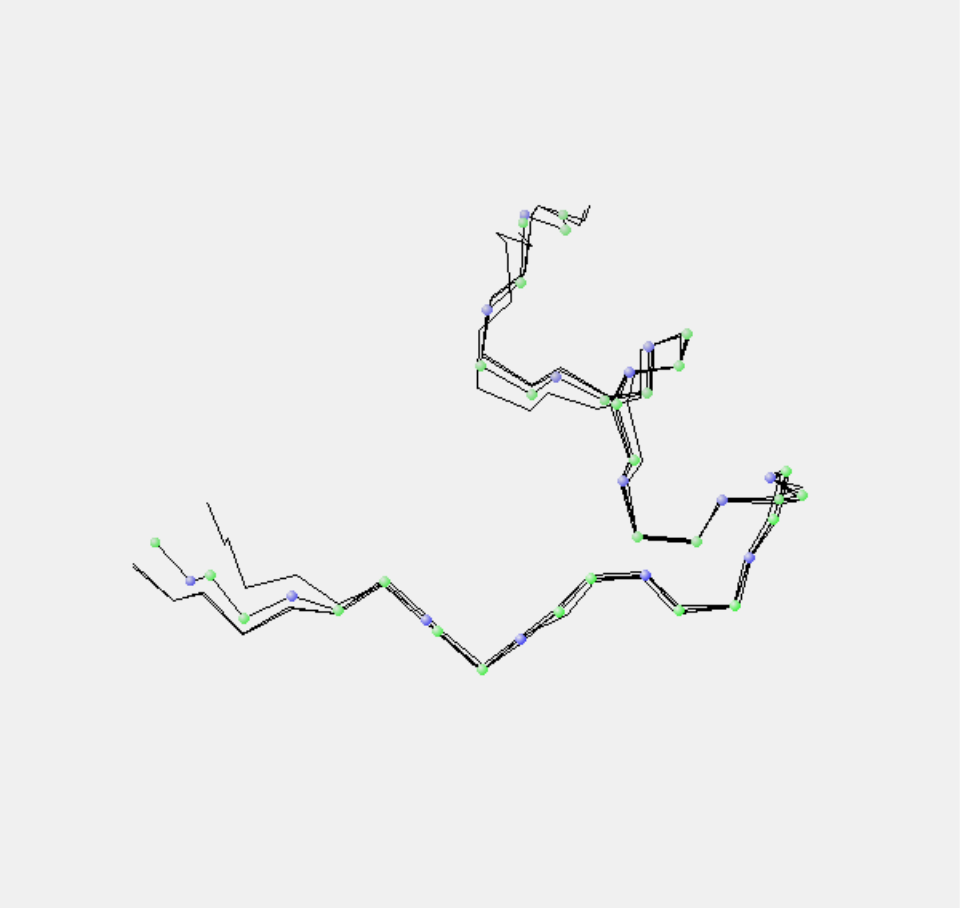

Figure 3: Clustered and aligned matches to pattern

The diagonal contribution to the quadratic form from which the errors associated with the mean cluster coordinates depend on contributions from all of the Δy_{li}. This is dominated by sums over $1/\Delta y_{li}^2$, so the error bars will often resemble the smallest RMSD for each of the clusters l. The only coupling with coordinate specific variations is through the rotation matrix cross terms. Since all of the Δy_{li}'s are assumed to be identical, the error bars show only slight variation, the error radius being about 0.3 $\mathring{\text{A}}$. Again, this reflects the variation of the mean, rather than sample-space variation.

A way to visualize this is to consider how much the mean could vary if any of the rigid cluster member segments were re-posed within the tolerances of the error bars. The rigidity of the member segments implies that variations that affect one part of the segment will be reflected uniformly in other parts as well given integration over all the available poses to estimate the variances.

Variations reported on the transformations indicate uncertainties in the angles of about 1.5°. Figure 3 shows the cluster in question. The aligned PDB segments are shown as simple stick figures; the average is marked by sticks with spheres. The sphere radii show the size of the estimated uncertainty in the cluster mean.

The more tolerant parameter set, $K = 8, K_{\max} = 2, R_{\max} = 2.5$ $\mathring{\text{A}}$. is perhaps more reasonable in that this is closer to the expected RMSD provided by x-ray crystallography. Compared to the first set's restrictive tolerances, the more relaxed tolerances accumulated more members in all of the clusters, and permitted the formation of another cluster, as seen in table 6. Generally the functional discrimination was also reduced.

Table 6: A.....A.A..VLA Cluster Members

RCSB Entry	Description
1aqn	Serine Protease (Subtilisin), Bacillus Amyloliquefaciens
1fiy	Lyase (Phosphoenolpyruvate Carboxylase), Escherichia Coli
1pjc	Oxidoreductase (L-Alanine Dehydrogenase), Phormidium Lapideum
2hsd	Oxidoreductase (Hydroxysteroid Dehydrogenase), Streptomyces Exfoliatus
2sic	Proteinase (Subtilisin), Bacillus Amyloliquefaciens
1arg	Transferase (Aspartate Aminotransferase), Escherichia Coli
1aup	Oxidoreductase (Nad-Specific Glutamate Dehydrogenase), Clostridium Symbiosum
1b3u	Scaffold Protein (Protein Phosphatase Pp2A), Homo Sapiens
1cov	Coxsackievirus Coat Protein, Coxsackievirus
1col	Antibacterial Protein (Colicin A), Escherichia Coli
1dpt	Cytokine (D-Dopachrome Tautomerase), Homo Sapiens
1flp	Oxygen Transport (Hemoglobin I), Lucina Pectinata
1mdt	Toxin (Monomeric Diphtheria Toxin), Corynebacterium Diphtheriae
1prc	Photosynthetic Reaction Center, Blastochloris Viridis
2hbg	Oxygen Transport (Hemoglobin), Glycera Dibranchiata
2trs	Lyase (Tryptophan Synthase), Salmonella Typhimurium
5ldh	Oxidoreductase (Lactate Dehydrogenase H4), Sus Scrofa
1b3n	Condensing Enzymes (Ketoacyl Acyl Carrier Protein Synthase), Escherichia Coli
1efp	Electron Transport (Electron Transfer Flavoprotein), Paracoccus Denitrificans
1b3n	Condensing Enzymes (Ketoacyl Acyl Carrier Protein Synthase 2), Escherichia Coli
1cdo	Oxidoreductase (Ch-Oh(D)-Nad(A))(Alcohol Dehydrogenase), Gadus Callarias
1deh	Oxidoreductase (Human β 1 Alcohol Dehydrogenase), Homo Sapiens
1efp	Electron Transport (Electron Transfer Flavoprotein), Paracoccus Denitrificans
1iea	Histocompatibility Antigen (Mhc Class II I-Ek), Mus Musculus
2ay1	Aminotransferase (Aromatic Amino Acid Aminotransferase), Paracoccus Denitrificans

Table 7: A.....A.A..VLA Average Coordinates

x	y	z	ΔY
20.1063	31.6247	16.3345	0.534537
19.1855	32.7291	15.8824	0.541982
17.9456	32.6423	16.3159	0.53143
16.9729	33.6246	15.9069	0.518602
17.3112	34.9742	16.4572	0.541988
17.6918	35.0067	17.7147	0.552098
18.0601	36.2597	18.3424	0.559707
19.1989	36.9219	17.6319	0.557689
20.1959	36.1304	17.3087	0.545851
21.364	36.6441	16.6388	0.528149
21.008	37.2466	15.3107	0.547005
20.1696	36.5668	14.5665	0.558134
19.7599	37.0679	13.271	0.557598
18.9723	38.3305	13.3904	0.557504
18.0785	38.3578	14.346	0.555484
17.2633	39.5356	14.5245	0.549708
18.1352	40.7152	14.8558	0.56434
19.0747	40.5151	15.7385	0.565801
19.9321	41.6058	16.1364	0.552415
20.7458	42.0627	14.9883	0.542332
21.1977	41.1289	14.2025	0.542142
21.9781	41.4342	13.0544	0.531104
21.1862	42.2247	12.0541	0.54599
19.8976	42.0852	12.0388	0.555307
18.9975	42.7591	11.1252	0.544145
18.4526	44.0542	11.6502	0.542197
18.8671	44.4566	12.7975	0.552627
18.3824	45.6708	13.3283	0.550638
17.5787	46.0122	14.1876	0.550019
16.9916	45.2914	14.6144	0.54976
16.1768	45.5973	15.4702	0.54687
16.5965	45.6554	16.8792	0.556911
16.1995	46.0798	17.5473	0.556049
16.4658	46.1468	18.925	0.557119
15.7823	45.2123	19.6964	0.559771
16.1674	44.8108	20.3285	0.554739
15.6006	43.9044	21.0929	0.555959
15.326	43.7843	22.3178	0.548894
14.3775	43.3626	22.5328	0.546673
13.9962	43.1927	23.6813	0.543744
13.8988	41.8981	24.2006	0.537171
14.6397	41.5908	24.9081	0.533829
14.6058	40.4158	25.4803	0.527968

The error bars show an error radius being about 0.55 Å. The modest size of the increase compared to the larger tolerances and greater variation was partly offset by the larger number of members in the cluster, since variances of averages scale inversely as the number of members. Figure 4 shows the larger estimated uncertainty in the mean segment as well as the broader range of variation between the segments that were clustered together at this tolerance. Variations reported on the transformations indicate uncertainties in the alignment angles of about 5.0°.

Table 8: A.....A.A..VLA Transformation Parameters

Parameter	Values		
	1aqn		
A	$1.00000 \pm 4.62147 \times 10^{-2}$	$-3.98688 \times 10^{-8} \pm 1.58513 \times 10^{-2}$	$5.18777 \times 10^{-8} \pm 2.48042 \times 10^{-2}$
	$3.98688 \times 10^{-8} \pm 4.62147 \times 10^{-2}$	$1.00000 \pm 1.58513 \times 10^{-2}$	$-2.12366 \times 10^{-7} \pm 2.48042 \times 10^{-2}$
	$-5.18777 \times 10^{-8} \pm 4.62147 \times 10^{-2}$	$2.12366 \times 10^{-7} \pm 1.58513 \times 10^{-2}$	$1.00000 \pm 2.48042 \times 10^{-2}$
$\vec{d}$	$7.52274 \times 10^{-7} \pm 6.77387 \times 10^{-2}$	$2.82677 \times 10^{-6} \pm 6.77387 \times 10^{-2}$	$-7.68530 \times 10^{-6} \pm 6.77387 \times 10^{-2}$
	1fiy		
A	$-9.04986 \times 10^{-1} \pm 1.08364 \times 10^{-1}$	$1.72486 \times 10^{-1} \pm 6.37527 \times 10^{-2}$	$3.88908 \times 10^{-1} \pm 3.92952 \times 10^{-2}$
	$3.09820 \times 10^{-1} \pm 1.08364 \times 10^{-1}$	$8.93673 \times 10^{-1} \pm 6.37527 \times 10^{-2}$	$3.24592 \times 10^{-1} \pm 3.92952 \times 10^{-2}$
	$-2.91569 \times 10^{-1} \pm 1.08364 \times 10^{-1}$	$4.14243 \times 10^{-1} \pm 6.37527 \times 10^{-2}$	$-8.62201 \times 10^{-1} \pm 3.92952 \times 10^{-2}$
$\vec{d}$	$4.69368 \times 10^{1} \pm 1.62657 \times 10^{-1}$	$-1.39607 \times 10^{1} \pm 1.62657 \times 10^{-1}$	$3.83412 \times 10^{1} \pm 1.62657 \times 10^{-1}$
	1pjc		
A	$9.64015 \times 10^{-1} \pm 9.85660 \times 10^{-2}$	$2.57251 \times 10^{-1} \pm 8.02482 \times 10^{-2}$	$6.70564 \times 10^{-2} \pm 4.93185 \times 10^{-2}$
	$-1.52075 \times 10^{-1} \pm 9.85660 \times 10^{-2}$	$7.40516 \times 10^{-1} \pm 8.02482 \times 10^{-2}$	$-6.54606 \times 10^{-1} \pm 4.93185 \times 10^{-2}$
	$-2.18054 \times 10^{-1} \pm 9.85660 \times 10^{-2}$	$6.20852 \times 10^{-1} \pm 8.02482 \times 10^{-2}$	$7.52990 \times 10^{-1} \pm 4.93185 \times 10^{-2}$
$\vec{d}$	$-9.46716 \pm 1.59653 \times 10^{-1}$	$4.39780 \times 10^{1} \pm 1.59653 \times 10^{-1}$	$1.21043 \times 10^{1} \pm 1.59653 \times 10^{-1}$
	2hsd		
A	$-8.23393 \times 10^{-1} \pm 3.83951 \times 10^{-2}$	$-4.12076 \times 10^{-1} \pm 3.87048 \times 10^{-2}$	$-3.90150 \times 10^{-1} \pm 5.09671 \times 10^{-2}$
	$-1.75395 \times 10^{-2} \pm 3.83951 \times 10^{-2}$	$7.05676 \times 10^{-1} \pm 3.87048 \times 10^{-2}$	$-7.08318 \times 10^{-1} \pm 5.09671 \times 10^{-2}$
	$5.67200 \times 10^{-1} \pm 3.83951 \times 10^{-2}$	$-5.76381 \times 10^{-1} \pm 3.87048 \times 10^{-2}$	$-5.88276 \times 10^{-1} \pm 5.09671 \times 10^{-2}$
$\vec{d}$	$8.14326 \times 10^{1} \pm 1.18084 \times 10^{-1}$	$7.26508 \times 10^{1} \pm 1.18084 \times 10^{-1}$	$-4.48893 \pm 1.18084 \times 10^{-1}$
	2sic		
A	$7.91177 \times 10^{-1} \pm 2.36960 \times 10^{-2}$	$-1.74914 \times 10^{-1} \pm 1.66504 \times 10^{-2}$	$5.86040 \times 10^{-1} \pm 4.24671 \times 10^{-2}$
	$-1.44341 \times 10^{-1} \pm 2.36960 \times 10^{-2}$	$8.77756 \times 10^{-1} \pm 1.66504 \times 10^{-2}$	$4.56848 \times 10^{-1} \pm 4.24671 \times 10^{-2}$
	$-5.94310 \times 10^{-1} \pm 2.36960 \times 10^{-2}$	$-4.46038 \times 10^{-1} \pm 1.66504 \times 10^{-2}$	$6.69213 \times 10^{-1} \pm 4.24671 \times 10^{-2}$
$\vec{d}$	$-1.35179 \times 10^{1} \pm 6.10813 \times 10^{-2}$	$5.96757 \pm 6.10813 \times 10^{-2}$	$3.95998 \times 10^{1} \pm 6.10813 \times 10^{-2}$

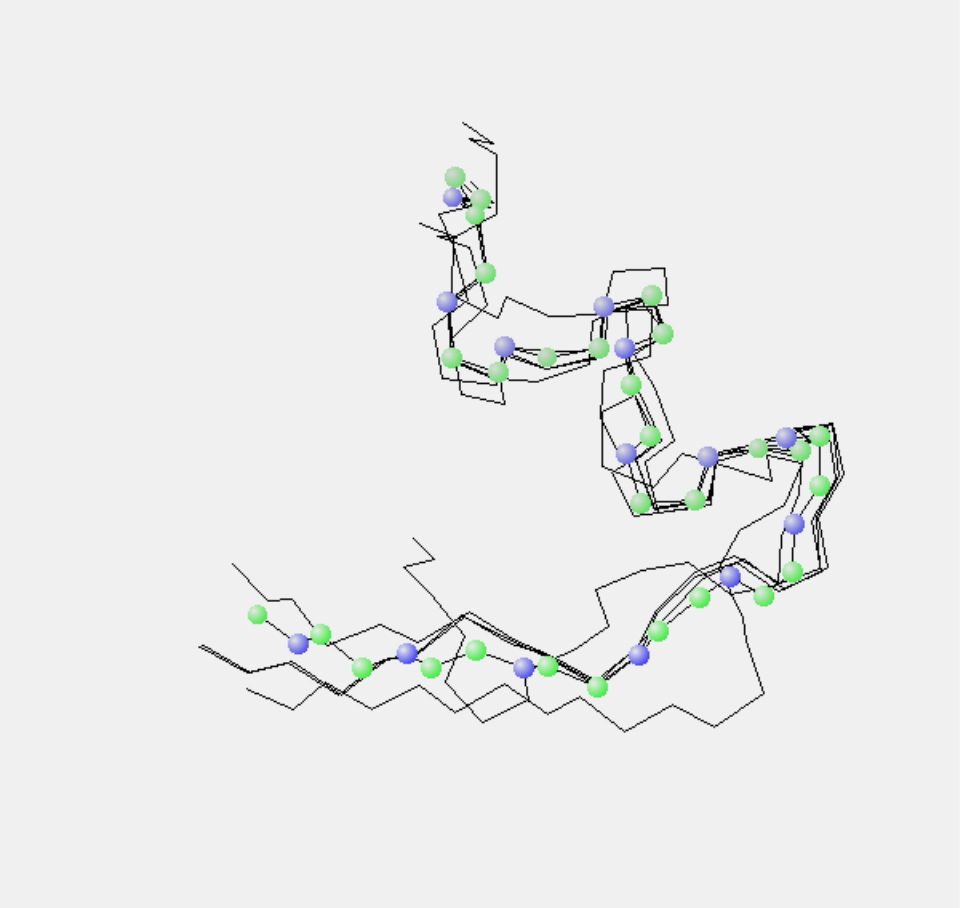

Figure 4: More tolerant clustering: clustered and aligned matches to pattern

4 Conclusions

This paper has presented techniques for analyzing the quality of clustered collections in terms of the statistical quality of the alignments, how to define a representative collection mean, and how to define a statistical characterization of similarity defined for polypeptide segments that match conserved segment patterns, as well as an exploration of the application to a single discovered pattern in a test case. If the cluster tolerance produce uncertainties in the coordinates that are large enough so that alignments between arbitrary shapes are attempted, the alignment angle uncertainties will be as large as the alignment angles themselves, which is a good indicator that the cluster tolerances are larger than discriminating structural features.

It was seen that cluster alignment quality was good for resolutions typical of X-ray diffraction studies. Further, increases in cluster tolerances increased cluster sizes and increased alignment uncertainties. Also, some functional discrimination was demonstrated, as well as the capability of recognizing conserved function and structure. As tolerances increased, that capability was somewhat diminished in the case of a single identified pattern. One notable exception followed from the limitation that given the small average seqlet span, W, there was little discrimination of function for the ubiquitous secondary structural α-helices.

Three dimensional alignment can be and has historically been very important in identifying conservation in proteins. Such information provides a way of recognizing which sites must be important to function, which has in turn

been a useful probe in studies of the mechanisms of enzyme function. Since databases are growing so rapidly, it becomes more advantageous to study alignments across the entire available databases, where it is prohibitively difficult to perform all the operations in the traditional way by hand. Methods involving the 3D clustering of correspondences between conserved sequence candidates provided by pattern discovery algoritms represent one automatable path towards the recognition of conserved protein structure as well as a possible tool for identifying function in the automation of annotation. Therefore, it is becomes important that the characterization of alignments should also contain information concerning the quality of the identification of the associations between candidate enzyme segments associated through a putative conservation.

A Least-Squares Estimate of Orientation Transformations

This section outlines the solution of the minimum χ^2 computation of an orientation transformation and its uncertainties expressed as covariances between the parameters and variances of the orientation transformation parameters. While the stability of the problem has been considered in the presence of noisy data,[16] a method for the propagation of errors must be defined.

Given that x_{ik} is the k'th component of the i'th atomic location vector to be transformed, y_{ij} is the j'th component of the i'th target location vector, a transformation would have the form

$$y_{ij} = \sum_k a_{jk} x_{ik} + d_j + e_{ij}, \tag{2}$$

where a_{jk} are the rotation matrix components of the orientation transformation, d_j are the displacement components of the orientation transformation, and e_{ij} is the statistically treated error in the j'th component of the i'th target atom location vector. If the deviations e_{ij} are independent and normally distributed about zero with a variance Δy_i^2, then each $e_{ij}^2/\Delta y_i^2$ contributes a single χ^2 degree of freedom. A sum of these deviations

$$\mathcal{E}^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j e_{ij}^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k a_{jk} x_{ik} + d_j - y_{ij} \right)^2 \tag{3}$$

may be constructed which measure the aggregate deviation of the alignment. If the e_{ij} are normally distributed, then $\mathcal{E}^2$ is χ^2 distributed with $3N$ degrees of freedom.

The deviation $\mathcal{E}^2$ measures the consistency of the data with a given set

of parameters α_{jk} and d_j. If $\mathcal{E}^2$ is χ^2, then the probability that the data is consistent with that particular alignment may be computed. It is then possible to ask what alignment is most consistent with the data, and how tolerant is that consistency to variations in the rotation and displacement parameters.

The solution will then proceed by computing the variation in the sum of squares of errors ($\mathcal{E}^2$) about a fixed set of parameters a_{jk} and d_j, and seeking values of those parameters consistent with minimizing the error [15][17] subject to the orthogonality constraints

$$\sum_j a_{jk} a_{jk'} = \delta_{kk'}. \tag{4}$$

The expansions are expressed by

$$d_j = D_j + \delta d_j, \tag{5}$$

$$a_{jk} = A_{jk} + \delta a_{jk}. \tag{6}$$

Solutions are then derived for the A_{jk} and D_j. These values will be represent best estimators of a_{jk} and d_j from the data. The covariances in the deviations $\delta a_{jk} = a_{jk} - A_{jk}$ and $\delta d_j = d_j - D_j$ that will be derived in the following represent variations to be expected in the estimators A_{jk} and D_j.

Covariances between rotation components, displacement components and cross-terms could be computed by extraction from the second order terms in the expansion of the error $\mathcal{E}^2$ about the best values, or by linearly propagating the covariance via expansions about the Δy_i variations in the y_{ij}'s. In the case of linearly propagated covariance, the covariance of $f(\xi_j)$ with $g(\xi_j)$ may be

computed in terms of $E(\delta f(\{\xi_j\})\delta g(\{\xi_j\}) = \sum_{jj'}(\partial f/\partial\xi_j)_0(\partial g/\partial\xi'_j)_0 E(\delta\xi_j\delta\xi_{j'})$. One of the solutions for the optimum rotation transformation, also called the orthogonal procrustes problem, involves a princpal components reduction of the rotation components[18] in a way that particularly ties the best-estimate rotation matrix to the eigenvectors of a matrix linear in the y_{ij}'s. While linear perturbations of eigensystems are well understood,[19] the definition of derivatives of degenerate eigenvectors with respect to linearly perturbing variations is not possible because the perturbation breakes the symmetry of the degenerate basis in a way that is strongly dependent on the perturbation.

However, even in the circumstances where linear propagation estimate of the covariances are poorly defined, the notion of error estimation is still sensible from the perspective of the variations of $\mathcal{E}^2$ about the optimum values. The use of the covariances deduced from $\mathcal{E}^2$ as employed in this presentation depends on the notion that the errors e_{ij} may be approximated by a normal distribution. This ammounts to the assertion that, at least as far as the estimation of error bars is concerned, a second order expansion of the distributions of the e_{ij}^2 is sufficient. In this situation, linear combinations of coordinate errors are also considered to be normally distributed, and therefore quadratic estimates of those errors may be considered to satisfy a multivariate normal distribution of the form $f(\vec{x}_j) \propto e^{-\mathcal{E}^2/2}$. If $\mathcal{E}^2$ is quadratic of the form $\mathcal{E}^2 = \mathcal{E}_0^2 + \sum_{jj'}\delta\xi_j C_{jj'}\delta\xi_{j'}$ for variations $\delta\xi_j$ in parameters ξ_j about some optimum value, then the covariance given by the expectation value $E(\delta\xi_j\delta\xi_{j'}$ is the inverse of the "Hessian matrix" specified by $C_{jj'}$.

Both the linear propagation technique and the evaluation of the inverse

of the Hessian produce essentially identical results in the case of simple linear regression. However, in this case of orthogonal constraints, these two approaches are not guaranteed to produce the same result, and particularly, the more common approach of simple linear propagation is subject to a loss of definition in the case of degeneracies. In this case, the $C_{jj'}$'s will have some zero eigenvalues associated with those components whose variances become undefined in the degenerate case such as occurs in problems with cylindrical symmetry.

The error may be expressed in terms of the above expansion as

$$\begin{aligned}
\mathcal{E}^2 &= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k A_{jk} x_{ik} + D_j - y_{ij} \right)^2 \\
&\quad + 2 \sum_{j'k'} \sum_i \frac{1}{\Delta y_i^2} \left(\sum_k A_{j'k} x_{ik} + D_{j'} - y_{ij'} \right) x_{ik'} \delta a_{j'k'} \\
&\quad + 2 \sum_{j'} \sum_i \frac{1}{\Delta y_i^2} \left(\sum_k A_{j'k} x_{ik} + D_{j'} - y_{ij'} \right) \delta d_{j'} \\
&\quad + \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k \delta a_{jk} x_{ik} + \delta d_j \right)^2
\end{aligned} \tag{7}$$

Those values of the parameters that minimize this error are those which eliminate linear dependences in δa_{jk} and δd_j. Since the δd_j's are not subject to externally imposed constraints, it follows that

$$D_j = \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1} \left[\left(\sum_i \frac{y_{ij}}{\Delta y_i^2} \right) - \sum_k A_{jk} \left(\sum_i \frac{x_{ik}}{\Delta y_i^2} \right) \right] \tag{8}$$

eliminates the contribution to the error from the δd_j's in the first order. Any deviation of the d_j's from these values produce only quadratic divergences in the error $\mathcal{E}^2$, implying those d_j's form a minimum in $\mathcal{E}^2$.

Substituting these back into the expression for $\mathcal{E}^2$ yields

$$\begin{aligned}\mathcal{E}^2 &= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right. \\ &\quad \left. + \sum_k \left[\xi_{ik} + \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1} \left(\sum_i \frac{x_{ik}}{\Delta y_i^2} \right) \right] \delta a_{jk} + \delta d_j \right\}^2 \end{aligned} \tag{9}$$

where

$$\xi_{ik} = x_{ik} - \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1} \left(\sum_i \frac{x_{ik}}{\Delta y_i^2} \right), \tag{10}$$

$$\eta_{ik} = y_{ik} - \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1} \left(\sum_i \frac{y_{ik}}{\Delta y_i^2} \right). \tag{11}$$

The coupling between displacement and rotation may be simplified by identifying

$$\delta d'_j = \delta d_j + \sum_k \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1} \left(\sum_i \frac{x_{ik}}{\Delta y_i^2} \right) \delta a_{jk}. \tag{12}$$

The error simplifies to become

$$\begin{aligned}\mathcal{E}^2 &= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} + \sum_k \xi_{ik}\delta a_{jk} + \delta d'_j \right\}^2 \\ &= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2 \\ &\quad + 2 \sum_i \frac{1}{\Delta y_i^2} \sum_{jk'} \delta a_{jk'} \xi_{ik'} \left(\sum_k A_{jk}\xi_{ik} - \eta_{ij} \right) \\ &\quad + 2 \sum_i \frac{1}{\Delta y_i^2} \sum_j \delta d'_j \left(\sum_k A_{jk}\xi_{ik} - \eta_{ij} \right) \\ &\quad + \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k \delta a_{jk}\xi_{ik} + \delta d'_j \right)^2 . \end{aligned}$$

The second cross term is identically zero since the sums over i of $\sum_i (1/\Delta y_i^2)\xi_{ik} = 0$ and $\sum_i (1/\Delta y_i^2)\eta_{ij} = 0$. The error simplifies further to

$$\mathcal{E}^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2$$

$$+2\sum_{jk'}\left[\sum_{k}A_{jk}\Xi_{kk'}-H_{k'j}\right]\delta a_{jk'}$$
$$+\sum_{i}\frac{1}{\Delta y_i^2}\sum_{j}\left(\sum_{k}\delta a_{jk}\xi_{ik}+\delta d'_j\right)^2, \qquad (13)$$

where

$$\Xi_{k'k} = \sum_{i}\frac{\xi_{ik'}\xi_{ik}}{\Delta y_i^2}, \qquad (14)$$

$$H_{k'j} = \sum_{i}\frac{\xi_{ik'}\eta_{ij}}{\Delta y_i^2}. \qquad (15)$$

These are the covariances of the coordinates about the means of the source coordinates (to be aligned) and target coordinates (to be aligned against) with the source coordinates weighted by the square of the coordinate uncertainties Δy_i^2.

The constraints imply that the coefficients of $\delta a_{jk'}$ need not be identically zero, in order for their contribution to cancel from the error. That constraint is

$$\sum_{j}a_{jk}a_{jk'}=\delta_{kk'}.$$

The second order expansion of this is

$$\delta_{kk'}=\delta_{kk'}+\sum_{j}A_{jk'}\delta a_{jk}+\sum_{j}A_{jk}\delta a_{jk'}+\sum_{j}\delta a_{jk}\delta a_{jk'},$$

or

$$\sum_{j}A_{jk'}\delta a_{jk}+\sum_{j}A_{jk}\delta a_{jk'}+\sum_{j}\delta a_{jk}\delta a_{jk'}=0. \qquad (16)$$

This represents six constraint equations satisfied by the nine δa_{jk}'s. Choose any $\lambda_{kk'}$ such that $\lambda_{kk'}=\lambda_{k'k}$ (symmetric). Then

$$2\sum_{kk'j}\lambda_{kk'}A_{jk}\delta a_{jk'}+\sum_{kk'j}\lambda_{kk'}\delta a_{jk}\delta a_{jk'}=0. \qquad (17)$$

In other words, if the coefficients of the $\delta a_{jk'}$ have the form $\sum_k \lambda_{kk'} A_{jk}$, then the contribution to $\mathcal{E}^2$ is quadratic in δa_{jk}. Thus if

$$\sum_k A_{jk}\Xi_{k'k} - H_{k'j} = \sum_k \lambda_{k'k} A_{jk} \tag{18}$$

or alternatively

$$\sum_k (\Xi_{k'k} - \lambda_{k'k}) A_{jk} = H_{k'j}, \tag{19}$$

then the linear contributions to $\mathcal{E}^2$ from the δa_{jk}'s are actually second order.

Recognizing the following vectors and matrix:

$$\vec{H}_{k'} = (H_{k'j}), \tag{20}$$

$$\vec{\Xi}_{k'} = (\Xi_{k'k}), \tag{21}$$

$$\vec{\lambda}_{k'} = (\lambda_{k'k}), \tag{22}$$

$$\mathbf{A} = (A_{jk}), \tag{23}$$

the above equation can be written

$$\mathbf{A} \cdot (\vec{\Xi}_{k'} - \vec{\lambda}_{k'}) = (\vec{\Xi}_{k'} - \vec{\lambda}_{k'}) \cdot \mathbf{A}^T = \vec{H}_{k'}.$$

Since $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{1}$, it follows that

$$(\vec{\Xi}_{k'} - \vec{\lambda}_{k'}) \cdot (\vec{\Xi}_k - \vec{\lambda}_k) = \vec{H}_{k'} \cdot \vec{H}_k, \tag{24}$$

which represents six equations for the six unknown $\lambda_{k'k}$'s. This can be interpreted by saying the $\vec{\Xi}_{k'}$ and $\vec{H}_{k'}$ form skewed basis vectors. The effect of the orthogonality constraint on the minimization of the least-square error will simply be the result of adding a symmetric $\lambda_{kk'}$ to the source basis so

that its basis vectors have the same lengths and relative angles as the target basis.

Rather than solving the above equations for the $\lambda_{kk'}$'s, an alternative solution for $\mathbf{A}$ may be obtained, whose results may be applied to obtain the $\lambda_{kk'}$'s. In this case, the problem is to minimize

$$\mathcal{E}^2 = \sum_{i=1}^{N} \sum_{j=1}^{3} \left(\frac{\sum_k A_{jk}\xi_{ik} - \eta_{ij}}{\Delta y_i} \right)^2$$

subject to the constraint that A_{jk} is a proper rotation. Solutions based on Singular Value Decomposition [20] that distinguish proper rotations from the space of general orthogonal matrices[18, 21] are well known, as are a number of other solutions.[22] Such a solution of $\mathbf{A}$ is numerically equivalent to the solution obtained by solving the six simultaneous equations for the $\lambda_{k'k}$'s. Since this is so, the $\lambda_{k'k}$'s may be obtained by using the $\mathbf{A}$ obtained by one of the other algorithms and applying it in

$$\vec{\lambda}_{k'} = \vec{\Xi}_{k'} - \mathbf{A}^{-1} \cdot \vec{H}_{k'}.$$

This evaluation of $\lambda_{k'k}$ may then be used in the error analysis that follows.

The form of the error at this point is

$$\begin{aligned}
\mathcal{E}^2 &= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2 + 2 \sum_{jk'} \left(\sum_k \Xi_{k'k} A_{jk} - H_{k'j} \right) \delta a_{jk'} \\
&\quad + \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k \delta a_{jk}\xi_{ik} + \delta d'_j \right)^2 \\
&= \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2 + \sum_{jkk'} \lambda_{kk'} \delta a_{jk} \delta a_{jk'} \\
&\quad + \sum_i \frac{1}{\Delta y_i^2} \sum_j \left(\sum_k \delta a_{jk}\xi_{ik} + \delta d'_j \right)^2 .
\end{aligned}$$

or

$$\mathcal{E}^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2 + \sum_{j'k'k''} \left(\Xi_{k'k''} + \lambda_{k'k''}\right) \delta a_{j'k''} \delta a_{j'k'} + \sum_j \left(\sum_i \frac{1}{\Delta y_i^2} \right) \delta d_j'^2 \qquad (25)$$

This is a quadratic form. The coefficients of this χ^2 quadratic form are generally related to the covariances in the δa_{jk} and $\delta d_j'$'s.[17] Since there are no cross terms including both $\delta d_j'$ and δa_{jk}, it follows that these variables are independently distributed. The covariance information may be extracted most simply by rewriting the collection

$$\delta\phi_l = (\delta a_{jk}) \qquad (26)$$

as a linear array. Then the quadratic form is

$$\mathcal{E}^2 = \mathcal{E}_0^2 + \sum_{ll'} \delta\phi_{l'} c_{ll'} \delta\phi_l + \sum_j \left(\sum_i \frac{1}{\Delta y_i^2} \right) \delta d_j'^2 \qquad (27)$$

where the $c_{ll'}$ form the symmetric matrix representing the coefficients in the quadratic form, and the residual error

$$\mathcal{E}_0^2 = \sum_i \frac{1}{\Delta y_i^2} \sum_j \left\{ \sum_k A_{jk}\xi_{ik} - \eta_{ij} \right\}^2 \qquad (28)$$

is the minimum error that could be obtained from the space of all possible a_{jk}'s and d_j's. Now, the $\delta\phi_l$'s may be rewritten to have the form

$$\delta\phi_l = \sum_j \varphi_{lj} \delta\Phi_j \qquad (29)$$

where

$$\sum_{l'} c_{ll'} \varphi_{l'j} = C_j \varphi_{lj}. \qquad (30)$$

In general, $\delta\phi_l$ has nine components. This implies there are $9 \times 10/2 = 45$ independent covariances. However, the nine δa_{jk}'s are subject to the six constraints of equation 15. The constraints do not reduce all 45 independent covariances to a number less than 9, so there can be nine non-zero principal values that result from the diagonalization $c_{ll'}$.

The φ_{lj} are orthogonal, and may be chosen to be orthonormal. Thus

$$\sum_l \varphi_{lj}\varphi_{lj'} = \delta_{jj'}, \tag{31}$$

$$\sum_j \varphi_{lj}\varphi_{l'j} = \delta_{ll'}. \tag{32}$$

It follows that the error has the form

$$\begin{aligned} \mathcal{E}^2 &= \mathcal{E}_0^2 + \sum_j C_j \delta\Phi_j^2 + \sum_j \left(\sum_i \frac{1}{\Delta y_i^2} \right) \delta d_j'^2 \\ &= \mathcal{E}_0^2 + \sum_{\substack{j \\ C_j \neq 0}} C_j \delta\Phi_j^2 + \sum_j \left(\sum_i \frac{1}{\Delta y_i^2} \right) \delta d_j'^2, \end{aligned} \tag{33}$$

If $C_j = 0$, $\delta\Phi_j$ makes no contribution to the error $\mathcal{E}^2$. The only variations that contribute to the error are those for which $C_j \neq 0$. Each of the quadratic contributions to $\mathcal{E}^2$ corresponds to a squared normally distributed variable. Therefore, each of the $\delta\Phi_j$'s with nonzero C_j's is normally distributed. The χ^2 distributed $\mathcal{E}^2$ variable is essentially composed of a multivariate normal distribution. Since $\mathcal{E}^2$ is diagonal with respect to the $\delta\Phi_j$'s, the $\delta\Phi_j$'s are independent of each other. Therefore, the distribution of $\delta\Phi_j$ varies like $Ae^{-C_j\delta\Phi_j^2/2}$, so that

$$E(\delta\Phi_j) = 0 \tag{34}$$

and

$$\mathrm{var}(\delta\Phi_j) = \frac{1}{C_j}. \tag{35}$$

Since there are no cross-terms, each of the $\delta\Phi_j$'s are independent. For the $\delta\Phi_j$'s, the covariance matrix is then

$$\langle \delta\Phi_j \delta\Phi_{j'} \rangle = \frac{\delta_{jj'}}{C_j}. \tag{36}$$

It follows that

$$\langle \delta\phi_l \delta\phi_{l'} \rangle = \sum_{\substack{jj' \\ C_j \neq 0}} \varphi_{lj}\varphi_{l'j'} \frac{\delta_{jj'}}{C_j}$$

or

$$\langle \delta\phi_l \delta\phi_{l'} \rangle = \sum_{\substack{j \\ C_j \neq 0}} \frac{\varphi_{lj}\varphi_{l'j}}{C_j} \tag{37}$$

This, together with the constraint information that any variations composed of

$$\delta\phi_l = \sum_{\substack{j \\ C_j = 0}} \varphi_{lj}\delta\Phi_j \tag{38}$$

produces no change in the error, gives error estimates for the computed parameters of the transformation.

Further, the $\delta d'_j$'s are independently distributed with variances

$$\mathrm{var}(\delta d'_j) = \left(\sum_i \frac{1}{\Delta y_i^2} \right)^{-1},$$

and a covariance of zero between different components.

While the $\delta d'_j$'s are independent of the δa_{jk}'s and of each other, it is not generally true of the δd_j's. Using the definition of the $\delta d'_j$'s, it follows that

$$\langle \delta d_j \delta d_{j'} \rangle \;=\; \langle \delta d'_j \delta d'_{j'} \rangle +$$

$$\sum_{kk'}\left[\left(\sum_i \frac{1}{\Delta y_i^2}\right)^{-2}\left(\sum_i \frac{x_{ik}}{\Delta y_i^2}\right)\left(\sum_i \frac{x_{ik'}}{\Delta y_i^2}\right)\right]\langle \delta a_{jk}\delta a_{j'k'}\rangle \quad (39)$$

$$\langle \delta d_j \delta a_{j'k}\rangle = -\sum_{k'}\left[\left(\sum_i \frac{1}{\delta y_i^2}\right)^{-1}\left(\sum_i \frac{x_{ik'}}{\delta y_i^2}\right)\right]\langle \delta a_{jk'}\delta a_{j'k}\rangle. \quad (40)$$

Since there are up to 9 degrees of freedom in the $\delta\Phi_j$'s, three degrees of freedom in the δd_k's, and there are $3N$ degrees of freedom in $\mathcal{E}^2$, it follows that there are up to $3(N-4)$ degrees of freedom remaining in $\mathcal{E}_0^2$. It is therefore possible to compute the probability that a smaller residual error composed of a sum of $3(N-4)$ normal variables could have occurred by chance (account should be taken of zero eigenvalue terms in the degenerate case). This may be applied as a measure of goodness of fit.

Alternatively, if the uncertainties in the coordinates Δy_i are unknown, an estimate may be determined in terms of a radius of uncertainty. Since each component contributes a Δy_i^2 to a squared radius, the expected radius of uncertainty R is defined by

$$R^2 = 3\Delta y_i^2, \quad (41)$$

where all the Δy_i's are assumed to be equal to each other. The expected error $\langle \mathcal{E}_0^2\rangle$ is just the number of degrees of freedom, so

$$\langle \mathcal{E}_0^2\rangle = 3(N-4) = \frac{3}{R^2}\sum_{ij}\left\{\sum_k A_{jk}\xi_{ik} - \eta_{ij}\right\}^2,$$

so that

$$R^2 = \frac{1}{N-4}\sum_{ij}\left\{\sum_k A_{jk}\xi_{ik} - \eta_{ij}\right\}^2.$$

This estimate of R may be employed to calculate the covariances between the δa_{jk}'s and the δd_j's that would be consistent with the noise in the data.

B Least-Squares Estimate of Average Alignments

This section outlines the solution of the minimum χ^2 computation of multiple alignments of molecular segments with some average segment, together with uncertainties expressed as covariances between the parameters and variances of the orientation transformation parameters as well as the coordinates of the average. The discussion closely parallels section 2, but there are a number of significant differences.

Given that x_{lik} is the k'th component of the i'th atomic location vector in the l'th molecular segment to be transformed, y_{ij} is the j'th component of the i'th target location vector in the *average* molecular segment, a transformation would have the form

$$y_{ij} = \sum_k a_{ljk} x_{lik} + d_{lj} + e_{lij}, \tag{42}$$

where a_{ljk} are the rotation matrix components of the orientation transformation for the l'th segment, d_{lj} are the displacement components of the orientation transformation for the l'th segment, and e_{lij} is the statistically treated deviation in the j'th component of the i'th target atom location vector in the l'th segment from the mean. If the deviations e_{lij} are independent and normally distributed about zero with a variance Δy_{li}^2, then each $e_{lij}^2/\Delta y_{li}^2$ contributes a single χ^2 degree of freedom. A sum of these deviations

$$\mathcal{E}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j e_{lij}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j \left(\sum_k a_{ljk} x_{lik} + d_{lj} - y_{ij} \right)^2 \tag{43}$$

may be constructed, which would be χ^2 distributed and which would mea-

sure the aggregate deviation of the alignment. The solution will proceed by computing the variation in the sum of squares of errors ($\mathcal{E}^2$) about a fixed set of parameters a_{ljk}, d_{lj}, and y_{lij}, and seeking values of those parameters consistent with minimizing the error [15][17] subject to the orthogonality constraints

$$\sum_j a_{ljk} a_{ljk'} = \delta_{kk'}. \tag{44}$$

The expansions are expressed by

$$d_{lj} = D_{lj} + \delta d_{lj}, \tag{45}$$

$$a_{ljk} = A_{ljk} + \delta a_{ljk}, \tag{46}$$

$$y_{jk} = Y_{jk} + \delta y_{jk}. \tag{47}$$

Solutions are then derived for the A_{ljk} and D_{lj}, and covariances between rotation components, displacement components and cross-terms will be computed by extraction from the second order terms in the expansion of the errors about the best values.

Note that since all of the components a_{ljk}, d_{lj}, and y_{jk} may be transformed by a global rotation and displacement leaving $\mathcal{E}^2$ invariant, there will be an ambiguity in the evaluation of these parameters. This ambiguity may be removed by choosing a_{Ljk} to be the identity matrix, and $d_{Lj} = 0$ for one of the molecular segments L without introducing any loss in generality.

The error may be expressed in terms of the above expansion as

$$\begin{aligned} \mathcal{E}^2 = & \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j \left(\sum_k A_{ljk} x_{lik} + D_{lj} - Y_{ij} \right)^2 \\ & + 2 \sum_{j'k'} \sum_{li} \frac{1}{\Delta y_{li}^2} \left(\sum_k A_{lj'k} x_{lik} + D_{j'} - Y_{ij'} \right) \delta a_{lj'k'} x_{lik'} \end{aligned}$$

$$+2\sum_{j'}\sum_{li}\frac{1}{\Delta y_{li}^2}\left(\sum_k A_{lj'k}x_{lik}+D_{lj'}-Y_{ij'}\right)\delta d_{lj'}$$
$$-2\sum_{j'}\sum_{li}\frac{1}{\Delta y_{li}^2}\left(\sum_k A_{lj'k}x_{lik}+D_{lj'}-Y_{ij'}\right)\delta y_{ij'}$$
$$+\sum_{li}\frac{1}{\Delta y_{li}^2}\sum_j\left(\sum_k \delta a_{ljk}x_{lik}+\delta d_{lj}-\delta y_{ij}\right)^2 \qquad (48)$$

Those values of the parameters that minimize this error are those which eliminate linear dependences in δa_{ljk}, δd_{lj}, and δy_{ij}. Since the δd_{lj}'s and δy_{ij}'s are not subject to externally imposed constraints, it follows that

$$D_{lj}=\left(\sum_i\frac{1}{\Delta y_{li}^2}\right)^{-1}\left[\left(\sum_i\frac{y_{ij}}{\Delta y_{li}^2}\right)-\sum_k A_{ljk}\left(\sum_i\frac{x_{lik}}{\Delta y_{li}^2}\right)\right] \qquad (49)$$

and

$$Y_{ij}=-\left(\sum_l\frac{1}{\Delta y_{li}^2}\right)^{-1}\left[\left(\sum_l\frac{y_{ij}}{\Delta y_{li}^2}\right)-\sum_{kl} A_{ljk}\frac{x_{lik}}{\Delta y_{li}^2}\right] \qquad (50)$$

eliminate the contribution to the error from the δd_{lj}'s and the δy_{ij}'s in the first order. Any deviation of the d_{lj}'s and y_{ij}'s from these values produce only quadratic divergences in the error $\mathcal{E}^2$, implying those d_{lj}'s and y_{ij}'s form a minimum in $\mathcal{E}^2$.

Substituting these back into the expression for $\mathcal{E}^2$ yields

$$\begin{aligned}\mathcal{E}^2 \;=\; & \sum_{li}\frac{1}{\Delta y_{li}^2}\sum_j\left\{\left(\sum_k A_{ljk}\xi_{lik}-\eta_{lij}\right)^2\right.\\ & +2\left(\sum_k A_{ljk}\xi_{lik}-\eta_{lij}\right)\sum_{k'}\delta a_{ljk'}x_{lik'}\\ & \left.+\left(\sum_k \delta a_{ljk}x_{lik}+\delta d_{lj}-\delta y_{ij}\right)^2\right\} \qquad (51)\end{aligned}$$

where

$$\xi_{lik} \;=\; x_{lik}-\left(\sum_i\frac{1}{\Delta y_{li}^2}\right)^{-1}\left(\sum_i\frac{x_{lik}}{\Delta y_{li}^2}\right), \qquad (52)$$

$$\eta_{lik} = Y_{ik} - \left(\sum_i \frac{1}{\Delta y_{li}^2}\right)^{-1} \left(\sum_i \frac{Y_{ik}}{\Delta y_{li}^2}\right). \quad (53)$$

The $\delta a_{ljk'} x_{lik'}$ term in the cross-term in $\mathcal{E}^2$ may be replaced to yield

$$\begin{aligned}\mathcal{E}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j \Bigg\{ & \left(\sum_k A_{ljk}\xi_{lik} - \eta_{lij}\right)^2 \\ & +2\left(\sum_k A_{ljk}\xi_{lik} - \eta_{lij}\right) \sum_{k'} \delta a_{ljk'} \left[\xi_{lik'} + \left(\sum_{i'} \frac{1}{\Delta y_{li'}^2}\right)^{-1} \sum_{i'} \frac{x_{li'k'}}{\Delta y_{li'}^2}\right] \\ & + \left(\sum_k \delta a_{ljk} x_{lik} + \delta d_{lj} - \delta y_{ij}\right)^2 \Bigg\}\end{aligned}$$

The term multiplying

$$\left(\sum_{i'} \frac{1}{\Delta y_{li'}^2}\right)^{-1} \sum_{i'} \frac{x_{li'k'}}{\Delta y_{li'}^2}$$

sums to zero from the same condition that cancels the linear coefficient of δd_{lj}. The error simplifies to

$$\begin{aligned}\mathcal{E}^2 = \sum_{li} \frac{1}{\Delta y_{li}^2} \sum_j \Bigg\{ & \left(\sum_k A_{ljk}\xi_{lik} - \eta_{lij}\right)^2 \\ & +2\left(\sum_k A_{ljk}\xi_{lik} - \eta_{lij}\right) \sum_{k'} \delta a_{ljk'} \xi_{lik'} \\ & + \left(\sum_k \delta a_{ljk} x_{lik} + \delta d_{lj} - \delta y_{ij}\right)^2 \Bigg\}.\end{aligned}$$

This is very similar to the situation in the simple alignment problem. Defining

$$\Xi_{lk'k} = \sum_i \frac{\xi_{lik'}\xi_{lik}}{\Delta y_{li}^2}, \quad (54)$$

$$H_{lk'j} = \sum_i \frac{\xi_{lik'}\eta_{ij}}{\Delta y_{li}^2}, \quad (55)$$

the contribution to $\mathcal{E}^2$ linear in $\delta a_{ljk'}$ is

$$2\sum_{ljk'}\left(\sum_{k} A_{ljk}\Xi_{lk'k} - H_{lk'j}\right)\delta a_{ljk'} = 0$$

at extremal A_{ljk}.

The constraints imply that the coefficients of the linear $\delta a_{ljk'}$ need not be identically zero, in order for their contribution to cancel from the error. That constraint is

$$\sum_{j} a_{ljk}a_{ljk'} = \delta_{kk'}.$$

The expansion of this about the extremal A_{ljk} is

$$\delta_{kk'} = \delta_{kk'} + \sum_{j} A_{ljk'}\delta a_{ljk} + \sum_{j} A_{ljk}\delta a_{ljk'} + \sum_{j}\delta a_{ljk}\delta a_{ljk'},$$

or

$$\sum_{j} A_{ljk'}\delta a_{ljk} + \sum_{j} A_{ljk}\delta a_{ljk'} + \sum_{j}\delta a_{ljk}\delta a_{ljk'} = 0. \tag{56}$$

Consider some symmetric

$$\lambda_{lkk'} = \lambda_{lk'k} \tag{57}$$

The resulting constraint between linear and quadratic δa_{ljk}'s is

$$2\sum_{kk'j} A_{ljk'}\delta a_{ljk}\lambda_{lkk'} + \sum_{jkk'}\lambda_{lkk'}\delta a_{ljk}\delta a_{ljk'} = 0. \tag{58}$$

The condition that A_{ljk} must satisfy to extremize $\mathcal{E}^2$ is the one for which the constraints conspire to make all the contributions from the linear δa_{ljk}'s quadratic. This occurs when

$$\sum_{k} A_{ljk}\Xi_{lkk'} - H_{lk'j} = \sum_{k}\lambda_{lkk'}A_{ljk}. \tag{59}$$

Again, this situation is identical to the orthogonal Procrustes problem, and the solution may be obtained by Singular Value Decomposition [20] that distinguish proper rotations from the space of general orthogonal matrices. [21] The error then becomes

$$\begin{aligned}\mathcal{E}^2 &= \sum_{li}\frac{1}{\Delta y_{li}^2}\sum_j\left\{\left(\sum_k A_{ljk}\xi_{lik}-\eta_{lij}\right)^2\right. \\ &\quad -\sum_{ljkk'}\lambda_{lkk'}\delta a_{ljk}\delta a_{ljk'} \\ &\quad \left.+\left(\sum_k \delta a_{ljk}x_{lik}+\delta d_{lj}-\delta y_{ij}\right)^2\right\}.\end{aligned}$$

This has been completely reduced to a quadratic form. As in the simple alignment case, the next step is to characterize the covariances via analysis of this quadratic form.

The direct coupling between displacement and rotation may be removed by identifying

$$\delta d'_{lj}=\delta d_{lj}+\sum_k\left(\sum_{i'}\frac{1}{\Delta y_{li'}^2}\right)^{-1}\left(\sum_i\frac{x_{lik}}{\Delta y_{li}^2}\right)\delta a_{ljk}. \tag{60}$$

The error simplifies to become (note that the coefficient of δa_{ljk} is now ξ_{lik} rather than x_{lik})

$$\begin{aligned}\mathcal{E}^2 &= \sum_{li}\frac{1}{\Delta y_{li}^2}\sum_j\left\{\left(\sum_k A_{ljk}\xi_{lik}-\eta_{lij}\right)^2\right. \\ &\quad -\sum_{ljkk'}\lambda_{lkk'}\delta a_{ljk}\delta a_{ljk'} \\ &\quad \left.+\left(\sum_k \delta a_{ljk}\xi_{lik}+\delta d'_{lj}-\delta y_{ij}\right)^2\right\}.\end{aligned}$$

The covariance $\mathrm{cov}(\delta a_{ljk},\delta d_{lj'})$ is expected to be zero since the sum over i in $\xi_{lik}\delta d'_{lj}$ cancels that coefficient of $\delta a_{ljk}\delta d_{lj'}$. However, it is not simple to

construct expressions that cancel the cross-terms of both δd_{lj} and δy_{ij} with each other and with δa_{ljk}. So, even though one of the cross-terms may be eliminated by the above transformation, the other cross-terms will induce non-zero covariances $\mathrm{cov}(\delta a_{ljk}, \delta d_{lj'})$.

The form of the error is now

$$\begin{aligned}
\mathcal{E}^2 &= \sum_{lij} \frac{1}{\Delta y_{li}^2} \left(\sum_k A_{ljk}\xi_{lik} - \eta_{lij} \right)^2 \\
&\quad + \sum_{ljkk'} \left(\sum_i \frac{\xi_{lik}\xi_{lik'}}{\Delta y_{li}^2} - \lambda_{lkk'} \right) \delta a_{ljk}\delta a_{ljk'} \\
&\quad + \sum_{lj} \left(\sum_i \frac{1}{\Delta y_{li}^2} \right) \delta d_{lj}'^2 \\
&\quad - 2 \sum_{lijk} \left(\frac{\xi_{lik}}{\Delta y_{li}^2} \right) \delta a_{ljk}\delta y_{ij} \\
&\quad - 2 \sum_{lij} \left(\frac{1}{\Delta y_{li}^2} \right) \delta d_{lj}'\delta y_{ij} \\
&\quad + \sum_{ij} \left(\sum_l \frac{1}{\Delta y_{li}^2} \right) \delta y_{ij}^2 .
\end{aligned} \tag{61}$$

This is a quadratic form. The coefficients of the quadratic form are generally related to the covariances in the δa_{ljk}'s, $\delta d_{lj}'$'s, and δy_{ij}'s.[17] The covariance information may be extracted most simply by rewriting the collection

$$\delta\phi_l = (\delta a_{ljk}, \delta d_{lj}', \delta y_{ij}) \tag{62}$$

as a linear array. Then the quadratic form is

$$\mathcal{E}^2 = \mathcal{E}_0^2 + \sum_{ll'} \delta\phi_{l'} c_{ll'} \delta\phi_l \tag{63}$$

where the $c_{ll'}$ form the symmetric matrix representing the coefficients in the

quadratic form, and the residual error

$$\mathcal{E}_0^2 = \sum_{lij} \frac{1}{\Delta y_{li}^2} \left(\sum_k A_{ljk} \xi_{lik} - \eta_{lij} \right)^2 \tag{64}$$

is the minimum error that could be obtained from the space of all possible a_{ljk}'s, d_{lj}'s, and y_{ij}'s. Now, the $\delta\phi_l$'s may be rewritten to have the form

$$\delta\phi_l = \sum_j \varphi_{lj} \delta\Phi_j \tag{65}$$

where

$$\sum_{l'} c_{ll'} \varphi_{l'j} = C_j \varphi_{lj}. \tag{66}$$

The φ_{lj} are orthogonal, and may be chosen to be orthonormal. Thus

$$\sum_l \varphi_{lj} \varphi_{lj'} = \delta_{jj'}, \tag{67}$$

$$\sum_j \varphi_{lj} \varphi_{l'j} = \delta_{ll'}. \tag{68}$$

It follows that the error has the form

$$\begin{aligned} \mathcal{E}^2 &= \mathcal{E}_0^2 + \sum_j C_j \delta\Phi_j^2 \\ &= \mathcal{E}_0^2 + \sum_{\substack{j \\ C_j \neq 0}} C_j \delta\Phi_j^2 \end{aligned} \tag{69}$$

If $C_j = 0$, $\delta\Phi_j$ makes no contribution to the error $\mathcal{E}^2$. The only variations that contribute to the error are those for which $C_j \neq 0$. Further, the constraints place extremum at a location in parameter space that is not necessarily the smallest value that the unconstrained quadratic form could have. This implies that some changes in the parameters could result in a smaller error. This manifests as occasional $C_j < 0$. Generally eigenvectors corresponding

to parameter variations with $C_j \leq 0$ are not consistent with the constraints, and should be excluded.

Each of the positive quadratic contributions to $\mathcal{E}^2$ corresponds to a squared Gaussian distributed variable. Therefore, each of the $\delta\Phi_j$'s with nonzero C_j's is Gaussian distributed. Further, the variance of the $\delta\Phi_j$'s are those values for which $\mathcal{E}^2$ would be increased by unity. In this case, it is clear

$$\mathrm{var}(\delta\Phi_j) = \frac{1}{C_j}. \tag{70}$$

Since there are no cross-terms, each of the $\delta\Phi_j$'s are independent. For the $\delta\Phi_j$'s, the covariance matrix is then

$$\langle \delta\Phi_j \delta\Phi_{j'} \rangle = \frac{\delta_{jj'}}{C_j}. \tag{71}$$

It follows that

$$\langle \delta\phi_l \delta\phi_{l'} \rangle = \sum_{\substack{jj' \\ C_j \neq 0}} \varphi_{lj}\varphi_{l'j'} \frac{\delta_{jj'}}{C_j}$$

or

$$\langle \delta\phi_l \delta\phi_{l'} \rangle = \sum_{\substack{j \\ C_j > 0}} \frac{\varphi_{lj}\varphi_{l'j}}{C_j} \tag{72}$$

This, together with the constraint information that any variations composed of

$$\delta\phi_l = \sum_{\substack{j \\ C_j \leq 0}} \varphi_{lj}\delta\Phi_j \tag{73}$$

produces no change in the error, gives error estimates for the computed parameters of the transformation.

Using the definition of the $\delta d'_j$'s, it follows that

$$\begin{aligned}
\langle \delta d_{lj} \delta d_{l'j'} \rangle &= \langle \delta d'_{lj} \delta d'_{l'j'} \rangle + \\
&\sum_{kk'} \left[\left(\sum_i \frac{1}{\Delta y_{li}^2} \right)^{-1} \left(\sum_i \frac{1}{\Delta y_{l'i}^2} \right)^{-1} \left(\sum_i \frac{x_{lik}}{\Delta y_{li}^2} \right) \left(\sum_i \frac{x_{l'ik'}}{\Delta y_{l'i}^2} \right) \right] \\
&\langle \delta a_{ljk} \delta a_{l'j'k'} \rangle \\
&- \sum_k \left[\left(\sum_i \frac{1}{\Delta y_{li}^2} \right)^{-1} \left(\sum_i \frac{x_{lik}}{\Delta y_{li}^2} \right) \right] \langle \delta a_{ljk} \delta d'_{l'j'} \rangle \\
&- \sum_k \left[\left(\sum_i \frac{1}{\Delta y_{l'i}^2} \right)^{-1} \left(\sum_i \frac{x_{l'ik}}{\Delta y_{l'i}^2} \right) \right] \langle \delta a_{l'jk} \delta d'_{lj'} \rangle \qquad (74) \\
\langle \delta d_{lj} \delta a_{l'j'k} \rangle &= \langle \delta d'_{lj} \delta a_{l'j'k} \rangle \\
&- \sum_{k'} \left[\left(\sum_i \frac{1}{\Delta y_{li}^2} \right)^{-1} \left(\sum_i \frac{x_{lik'}}{\Delta y_{li}^2} \right) \right] \langle \delta a_{ljk'} \delta a_{l'j'k} \rangle. \qquad (75)
\end{aligned}$$

There are 6 independent $\delta\Phi_j$'s from each of the G segment matrices δa_{ljk}, 3 independent $\delta\Phi_j$'s from each of the G segment displacements δd_{lj}, and 3N independent $\delta\Phi_j$'s from the δy_{ij}'s, for a total number of $\delta\Phi_l$'s to be $9G+3N$, which is the maximum possible number of non-zero C_j's. The total number of contributing terms to $\mathcal{E}^2$ is $3NG$. So, the number of degrees of freedom in the residual error $\mathcal{E}_0^2$ is $3NG - n_0$, where n_0 is the number of non-zero C_j's.

Alternatively, if the uncertainties in the coordinates Δy_i are unknown, an estimate may be determined in terms of a radius of uncertainty. Since each component contributes a Δy_i^2 to a squared radius, the expected radius of uncertainty R is defined by

$$R^2 = 3\Delta y_i^2, \qquad (76)$$

where all the Δy_i's are assumed to be equal to each other. The expected

error $\langle \mathcal{E}_0^2 \rangle$ is just the number of degrees of freedom, so

$$\langle \mathcal{E}_0^2 \rangle = 3NG - n_0 = \sum_{lij} \frac{3}{R^2} \left(\sum_k A_{ljk} \xi_{lik} - \eta_{lij} \right)^2 ,$$

so that

$$R^2 = \frac{3}{3NG - n_0} \sum_{lij} \left(\sum_k A_{ljk} \xi_{lik} - \eta_{lij} \right)^2 .$$

This estimate of R may be employed to calculate the covariances between the δa_{ljk}'s, the δd_j's, and the δy_{ij}'s that would be consistent with the noise in the data.

C Modified Jarvis-Patrick Clustering Model

Once a pattern has been identified, all of the coordinates of the C – C_α – N atoms in the peptide bases matching the pattern are extracted. RMS distances between the best alignments of each pair of the pattern matches are computed. The computation of a best pairwise alignment and error analysis of the alignment estimates are described in Appendix 1. Clustering then follows the following algorithm:

1. Sequentially index each element of the array of matches. These sequential indices are the initial cluster identification numbers.

2. Construct nearest neighbor lists for each array element using the above described RMS distances.

3. For each pair of array entries i and j $(i < i)$, if all of the following conditions are met:

   - i is in the top K of j's nearest neighbor list;
   - j is in the top K of i's nearest neighbor list;
   - i and j must have $K_{\min}$ out of the top K entries in their nearest neighbor lists in common;
   - the RMS distance between i and j must be less than some $R_{\max}$.

   the j and i are considered to be valid cluster neighbors, and j is assigned the same cluster number that i has.

4. Once all the common cluster memberships are determined, the computation of an average backbone representing the cluster together with its error analysis, as described in Appendix 2, is computed.

The clustering algorithm requires the specification of K (the number of neighbors in the neighbor lists to consider), $K_{\min}$ (the minimum number of nearest neighbors out of K that must agree between valid cluster neighbors), and $R_{\max}$ (the maximum RMS distance between valid cluster members). This algorithm is essentially identical to the Jarvis-Patrick algorithm [2] except for the additional $R_{\max}$ constraint.

References

[1] Rigoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida, "Dictionary Building Via Unsupervised Hierarchical Motif Discovery In The Sequence Space of Natural Proteins," *Journal of Proteins: Structure, Function and Genetics,* **37(2)**, 264-277(1999).

[2] Jarvis, R. A., and Patrick, E. A.; "Clustering using a Similarity Measure Based on Shared Nearest Neighbors," *IEEE Transactions on Computers*, **C-22**, 1025-1034(1973).

[3] Jonassen, I. E., "Methods for Discovering Conserved patterns in Protein Sequences and Structures," in *Bioinformatics: Sequence, Structure, and Databanks*, Higgins, D. and Taylor, W., eds, Oxford University Press, NY, 143-166(2000).

[4] Jonassen, I. E., Grindhaug, S. H., Taylor W. R., "Searching the Protein Structure Databank with Weak Sequence Patterns and Structural Constraints," *J. Mol. Biol.*, **304**, 597-617(2000).

[5] Holm, L. and Sander, C., "Protein Structure Comparison by Alignment of Distance Matrices," *J. Mol. Biol.* **233**, 123-138(1993).

[6] Holm, L. and Sander, C., "Parser for Protein Folding Units," *Proteins*, **19**256-268(1994), Holm, L. and Sander, C., "Mapping the Protein Universe," *Science*, **273**, 595-603(1996).

[7] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C., "SCOP: A structural Classification of Proteins Database for the Investigation of Sequences and Structures," *J. Mol. Biol.* **247**, 536-540(1995).

[8] Orengo, C. A., Brown, N¿ P., Taylor, W. R., "Fast Structure Alignment for Protein Databank Searching," *Proteins*, **14**, 139-167(1992).

[9] Chothia, C., "One Thousand Families for the Molecular Biologist," *Nature*, **357**, 543-544(1992).

[10] Wang, Z-X, "How Many Fold Types of Protein Are There in Nature?" *Proteins: Structure, Function, and Genetics,* **26**, 186-191(1996).

[11] Govindarajan, S., Recabarren, R., Goldstein, R. A., "Estimating the Total Number of Protein Folds," *Proteins: Structure, Function, and Genetics,* **35**, 408-414(1999).

[12] Chapman, D., "The Measurement of Molecular Diversity: A Three-Dimensional Approach," *J. Computer-Aided Molecular Design*, **10**, 501-512(1996).

[13] Hua, XY and Raleigh, D. "On the Global Architecture of Initiation Factor IF3: A Comparative Study of the Linker Regions from the Escherichia coli Protein and the Bacillus stearothermophilus Protein," *J. Mol. Biol.* **278**, 871-878(1998).

[14] Escalier, V., Potheir, J., Soldano, H., Viari, A.; "Pairwise and Multiple Identification of Three-Dimensional Common Substructures in Proteins," *J. Comp. Biol.* **5**, 41-56(1998).

[15] Freund, J. E.; *"Mathematical Statistics, 5th ed."* Prentice Hall, Upper Saddle River, NJ, 1992

[16] D. Goryn, "On the Estimation of Rigid Body Rotation from Noisy Data," *IEEE Pattern Anal. and Machine Intell.* **17**, 1219-1220(1995)

[17] Bevington, P. R.; *"Data Reduction and Error Analysis for the Physical Sciences"*, McGraw-Hill, NY, 1969.

[18] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares Fitting of Two 3-D Point Sets," *IEEE Pattern Anal. and Machine Intell.* **9**, 698-700(1987)

[19] R. M. Haralick, "Principles of Covariance Propagation," *Proceedings of SPIE - The International Association for Optical Engineering*, **3811**, 2-19(1999).

[20] Press, W. H.; Teukolski, S. A.; Vetterling, W. T.; Flannery, B. P.; *"Numerical Recipes in C, 2nd ed,"* Cambridge University Press, NY, 1992.

[21] K. Kanatani, "Analysis of 3-D Rotation Fitting," *IEEE Pattern Anal. and Machine Intell.* **16**, 543-549(1994)

[22] D. W. Eggert, A. Lorusso, and R. B. Fisher, "Estimating 3-D Rigid Body Transformations: a Comparison of Four Major Algorithms," *Machine Vision and Applications*, **9**, 272-290(1997)

Figures

1. Backbones of all matches to $A.....A.A..VLA$.

2. Backbones of entries in cluster 1 from matches to $A.....A.A..VLA$, unaligned.

3. Backbones of entries in cluster 1 from matches to $A.....A.A..VLA$, aligned. with $R_{\max} = 1.0$ Å.

4. Backbones of entries in cluster 1 from matches to $A.....A.A..VLA$, aligned with $R_{\max} = 2.5$ Å.