Covariation of residues in the homeodomain sequence family

NEIL D. CLARKE

Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, Baltimore, Maryland 21205

(RECEIVED July 12, 1995; ACCEPTED August 22, 1995)

Abstract

Homeodomains are 60 amino acid DNA binding domains found in numerous eukaryotic transcription factors. The homeodomain family is a useful system for studying sequence–structure relationships because several hundred sequences are known and the structures of several homeodomains have been determined. Covariation of amino acid residues in the homeodomain family has been investigated to see whether strongly covariant residue pairs can be understood in terms of the structure and function of these domains. Among 16 strongly covariant pairs examined, 2 are explained by the ability to form salt bridges, and 9 appear related to the DNA binding function of the proteins. For the remaining 5 pairs, the rationale for covariance remains unclear and the likelihood of artifactual correlations is discussed in the context of experimental and evolutionary biases in the selection of sequences. No significant correlation was found between covariance and structural proximity in the hydrophobic core.

Keywords: covariation; evolution; homeodomains; sequences; structure–function

Sequence families contain more information than single sequences, and the rapid growth in the number and size of protein sequence families has engendered considerable interest in learning how to use this extra information. Homeodomains are an interesting example of a large and diverse sequence family that can provide a useful test system for studies of this type. There are now several hundred sequences of this very small and simple motif, and the structures of five homeodomains (Antp, en, ftz, MAT α 2, and Oct1) have been determined (Qian et al., 1989; Kissinger et al., 1990; Wolberger et al., 1991; Billeter et al., 1993; Kappen et al., 1993; Clarke et al., 1994; Klemm et al., 1994; Qian et al., 1994). The structures of the homeodomains are closely related. For example, equivalent α -carbons for residues 9–58 of the homeodomains from *Drosophila* engrailed and yeast MAT α 2 can be superimposed with an RMS deviation of 1.0 Å (Wolberger et al., 1991). The sequences of these two domains are 27% identical, somewhat less than the median identity (40%) for all pairs of sequences among the 263 used in this study. However, the degree of sequence divergence is highly nonuniform. There are clear sequence subfamilies, some containing large numbers of closely related sequences, others having only one or a few (Laughon, 1991; Kappen et al., 1993). In some cases, a hierarchy or nesting of subfamilies is

evident, as well. Skewed distributions of this type are an inevitable feature of biologically derived sequences and have important, if poorly understood, implications for analysis.

An amino acid that contributes to the stability or function of a protein in one sequence context may be dispensable or even unfavorable in another. In evolutionary terms, the “fitness” of an amino acid at some position in a protein depends on the amino acids at other positions. An extreme example of this would be an aspartic acid at a sequence position that is buried in the native fold of the protein. If the native fold allowed formation of a salt-bridge between the aspartic acid and a buried lysine, then aspartic acid at this position would be expected to have a high fitness level relative to other amino acids. However, replacement of the lysine with a hydrophobic amino acid would drastically decrease the fitness of the aspartate. The fitness of amino acids at other residue positions would be expected to change as well, but probably not as much as that of the aspartate because structural or functional links between the lysine and other residues are likely to be much less direct.

The fact that an amino acid change at one position affects the fitness of amino acids at other positions in a manner that depends on the structural and functional “links” between the residues suggests that evolutionarily related sequences should contain the vestiges of these effects in the form of correlated (covariant) residue pairs. A number of recent papers have examined the relationship between residue correlations and structure (Gobel et al., 1994; Neher, 1994; Shindyalov et al., 1994; Taylor & Hatrick, 1994). However, as noted by these authors and

Reprint requests: Neil D. Clarke, 708 WBSB, Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, 725 N. Wolfe Street, Baltimore, Maryland 21205; e-mail: neil.clarke@qmail.bs.jhu.edu.

others, apparent correlations can also arise from biases in the sequence set.

The example of the buried salt bridge serves to illustrate a distinction made in this paper between residue pairs and amino acid pairs. The term "residue pair" will be used to refer to a pair of amino acid residue positions (residue numbers). "Amino acid pair" will be used to refer to a particular pair of amino acids found at a residue pair. For example, if D is the amino acid at residue i and K is the amino acid at residue j , the residue pair is i - j and the amino acid pair at i - j is D-K. Different amino acid pairs can contribute very differently to the overall correlation at a residue pair. In this example, oppositely charged amino acid pairs are expected to contribute strongly to the residue pair correlation because each amino acid has a significant effect on the fitness of the other amino acid in the pair. However, if there are sequences in which the salt bridge has been replaced by a pair of hydrophobic residues, such amino acid pairs might be expected to contribute less to the residue pair correlation. This is because substitution of one hydrophobic amino acid for another at position i will generally have a less dramatic effect on the fitness of a second hydrophobic amino acid at position j than is the case for the buried salt bridge.

Before proceeding to the analysis of homeodomain sequences, it is instructive to consider what our expectations would be for an ideal set of sequences and how real sequence sets differ from the ideal. Suppose, for example, that we knew every possible amino acid sequence that could adopt the homeodomain fold. In this simple case, any covariation in a residue pair has to have structural significance because there is no possibility of statistical artifacts of any type. Because the ability to fold is the only criterion for the inclusion of a sequence in this set, there must be some kind of linkage in terms of protein folding between residues that are covariant.

If we now add functional criteria (e.g., DNA binding) to select a subset of the folding-competent sequences, the situation becomes slightly more complicated. Some residue pairs may now be covariant because they (or at least a subset of amino acid pairs at those residues) work together in binding a particular DNA sequence. In addition, there may also be interactions between the selective pressures for folding and for DNA binding. For example, subtle structural differences among the sequences that meet the folding criterion could influence the functional fitness of different combinations of DNA binding residues. As a result, correlated residue pairs could exist in which one residue plays a direct role in DNA binding while the second has a subtle effect on the structure of the protein. Despite the addition of this second selection criterion, covariation still implies structural or functional linkage provided (1) all sequences are subject to the same evolutionary selection criteria and (2) the set of sequences is sufficiently large and representative.

If it is not the case that all sequences are subject to the same selective pressures, then there is no longer necessarily a connection between covariation of a residue pair and linkage in a structural or functional sense. Suppose, for example, that a homeodomain that binds to a particular DNA sequence comes under selective pressure to also bind another protein. Suppose further that this homeodomain, and its evolutionary descendants, undergo further sequence divergence subject to selective pressure for protein folding, DNA binding, and protein binding. If this subfamily of sequences is combined with other sequences and analyzed for covariation, then there will be an

apparent correlation between residues critical for site-specific DNA binding and those critical for binding the second protein.

This artifactual covariance between DNA binding residues and protein binding residues would disappear, however, if the sequences that bound the second protein further diverged to the point where the distribution of DNA binding specificities was the same as for those homeodomains that do not bind the second protein. If a sufficiently large number of such sequences were known, there would be no connection between variation at the DNA binding residues and variation at residues that bind the second protein. Artifactual covariance arises, therefore, because of different selective pressures and because the evolutionary and experimental sampling of sequences is incomplete and unrepresentative.

One problem with obtaining representative sequence data is that the sampling of sequences is not phylogenetically uniform. For example, there are a disproportionately large number of homeodomain sequences derived from *Drosophila* and mammals. In the case of the homeodomains, however, this is probably not serious because subfamilies of homeodomains tend to be phylogenetically widespread, suggesting a very early evolutionary origin and divergence for this motif; this minimizes the impact of selective phylogenetic sampling (Kappen et al., 1993). A more serious source of bias is that many of the known sequences of homeodomains are from clones obtained on the basis of sequence similarity to existing homeodomain sequences or have been identified on the basis of close sequence similarity. This can be expected to lead to a more pronounced clustering of closely related sequences into subfamilies than is the case for the true set of biologically extant sequences.

A less obvious source of sample bias is that the subset of possible homeodomain sequences that have been sampled in the course of evolution is itself a function of the evolutionary history of the domain. In a sense, sampling the "sequence space" of a structure is rather similar to sampling conformational space in protein folding. Local minima in sequence space are likely to exist, and some evolutionary pathways for sequence divergence are almost certainly more likely than others. The paths and branches taken depend on the sequence of the ancestral homeodomain and on the sequences that happen to exist at various branch points.

It is not clear what the most appropriate way to handle sequence bias is. One approach is to attempt to fill in missing evolutionary information by constructing a parsimonious sequence similarity tree and then use the tree to infer the evolutionary occurrence of "simultaneous" substitutions at two positions (Shindyalov et al., 1994). A second approach is to construct a subset of sequences such that each sequence differs from all other sequences by some threshold of amino acid identity or similarity (Neher, 1994; Taylor & Hatrick, 1994). A variation of this approach is to down-weight the contributions of each sequence depending on its similarity to other sequences in the set (Gobel et al., 1994).

Unfortunately, we cannot know how the set of known sequences differs from the possible sequence set, and elimination or downweighting of sequences necessarily affects the information content of those sequences. Although large numbers of closely related sequences result in biases in the observed correlations, they also help provide statistical significance to those observations. The approach adopted here is to use a complete set of sequences as well as reduced sequence sets that are con-

structed by iteratively eliminating those sequences that contribute most to the bias.

Results

Sequence sets used in this analysis

The analysis reported here was performed on three sets of homeodomain sequences, the second and third sets being subsets of the first. The first set consists of the 263 unique, full-length (60 amino acid) homeodomain sequences listed in a recently published compilation and analysis (Kappen et al., 1993). The published list contains a larger number of sequences, but where there are identical sequences, only one was retained for this study. Sequences that are shorter at either end than the conventionally defined 60 amino acids also were eliminated. The alignment of several sequences (MAT α 2, prl, Zn1, ZMH1, and ZMH2) was corrected from that given in the published compilation due to the presence of a short insertion between helices 1 and 2 in these domains. This insertion was predicted earlier on the basis of sequence comparisons and was confirmed by the structure determination of MAT α 2 (Laughon, 1991; Wolberger et al., 1991). The second and third sequence sets were constructed in turn from the first by the elimination of those sequences that contributed most to the highly correlated pairs identified in the previous set. For example, the most highly covariant residue pairs identified in the complete sequence set were 19-30, 28-54, and 44-46. The amino acid pairs that contributed most strongly to the covariance of these residue pairs were E19-R30, R28-M54, and Q44-K46, respectively. Ninety of the 263 unique sequences contain the amino acids E19, R30, R28, M54, Q44, and K46. These 90 sequences were eliminated to form the second sequence set, which therefore contains 173 sequences. The third set was derived in an analogous manner. The most highly covariant residue pairs in the second sequence set are 44-46, 46-47, and 46-50, and the amino acid pairs that contribute most to the covariance of these pairs are Q44-K46, K46-I47, and K46-Q50, respectively. Elimination of all sequences in the second set that contain Q44, K46, I47, and Q50 leaves 87 sequences remaining in the third set.

Any procedure used to construct or weight sequence sets necessarily alters the results and there are no strong theoretical grounds for favoring one procedure over another. However, the procedure used here has the merit that the criterion used for elimination of sequences is based on covariance (rather than simple conservation) and it produces sequence subsets in which the unavoidable biases that exist in the original set are at least changed in a sensible way. Perhaps the strongest justification (which could also be true for other schemes) is that additional highly correlated residues pairs that are both interesting and rationalizable were found in the derived sequence sets.

Figure 1 shows the effect of the iterative elimination of sequences on the overall complexity of the sequence set. The number of identical residues was determined for all possible pairs of sequences. The figure shows, for each sequence set, the relative number of sequence pairs having a given number of identical residues. There are clearly fewer closely related sequence pairs in the 87 sequence set than in the 263 sequence set, suggesting that the average complexity of the set has increased. The other notable feature in this figure is that the distribution of sequence pair identities does not fit a normal distribution. Peaks in the distribution are a consequence of the clustering of se-

Fig. 1. Distribution of pairwise sequence identities for each of the three sequence sets. The number of sequence pairs for each sequence set ($263 \times 262/2$; $173 \times 172/2$; $87 \times 86/2$) have been normalized.

quences into subfamilies and, although this clustering persists in the smallest sequence set, it is somewhat less pronounced.

Definition of covariance or "mutual information"

Here, the degree to which amino acids at residue pair i, j show covariance is defined as

$$\sum_n P_{a_i, a_j} \log(P_{a_i, a_j} / (P_{a_i} P_{a_j})),$$

where the summation is over all sequences. a_i and a_j are the amino acids found at residues i and j , respectively, in sequence n . P_{a_i} is the probability that amino acid a_i would be found at residue i , and is considered equivalent to the observed frequency with which that amino acid is found at residue i among all sequences in the set. P_{a_j} is the probability (observed frequency) of finding amino acid a_j at residue j , and P_{a_i, a_j} is the probability (observed frequency) of finding both a_i at residue i and a_j at residue j in the same sequence. Thus, $P_{a_i} P_{a_j}$ (the product of the individual probabilities for a_i and a_j) is the expected frequency with which amino acid pair a_i, a_j will be found at residue pair i, j if they are independent of one another, whereas P_{a_i, a_j} is the actual observed frequency for the amino acid pair. This definition of covariance is similar, but not identical, to that used by Korber et al. (1993) in their treatment of covariation in the HIV envelope protein, and the reader is referred to their paper for a useful introduction to the information theoretic basis for metrics of this type. The use of the term "mutual information" reflects the idea that covariance is a measure of how much information about the identity of one residue in a pair is contained in the identity of the second.

The definition given above for covariance (or mutual information) can be written equivalently as

$$\sum_{a_i, a_j} (P_{a_i, a_j})^2 \log(P_{a_i, a_j} / (P_{a_i} P_{a_j})).$$

Here, the summation is over all amino acid pair types rather than over all sequences, and in this form it can be seen to differ from the more theoretically rigorous definition of Korber et al. (1993)

in the squaring of P_{a_i, a_j} . The definition used here has the effect of weighting information-rich amino acid pairs by the number of times they are observed. This weighting was done because it was found empirically to give many more covariant residue pairs that were easily rationalized in structural or functional terms than was the case for those identified without weighting.

Furthermore, in the absence of weighting, the residue pairs identified as being highly covariant tended to be characterized by large numbers of amino acid pairs found at low frequency. The influence of rare pairs on the calculation of covariance was of concern because the sequence set contains some highly diverged sequences that may be inappropriately aligned. A comparison of results obtained with the sequence set that contains five misaligned sequences with those obtained using the corrected sequence set indicates that the weighted covariation value is relatively insensitive to misalignments: the top 10 residue pairs found using the incorrect set were the same 10 found with the correctly aligned sequences except that the rank order of 2 pairs adjacent in the list was reversed.

Covariance of residue pairs in the homeodomain sequence family

Figure 2 is a graphical representation of the mutual information value for each residue pair. Each panel in the figure corresponds to a different sequence set. The values have been normalized within each sequence set such that the maximum and minimum values in each matrix correspond to the extremes of the color bar shown in Figure 2A. There is no directionality to the definition of mutual information, so residue pair (i, j) is equivalent to (j, i) .

The matrices are dominated by blue and purple elements (low mutual information), indicating that the median value for residue pair covariance is considerably lower than the mid-point of the extreme values; the distribution of values is considered in more detail below. A few residue pairs are unusually covariant, as evidenced by the yellow to red elements. Note that there are purple stripes due to the very low mutual information value for all pairs of residues involving residues 48, 49, 51, and 53. The reason for this is that residue 48 is tryptophan in all 263 sequences, and residues 49, 51, and 53 are identical in more than 97% of all sequences. Consequently, the identity of the amino acid at each of these positions cannot contain much information (none in the case of Trp 48) regarding the identity of amino acids at other positions. This illustrates an important distinction between residues that show significant covariation with other residues and those that are highly conserved. Although both may play critical roles in the structure or function of a protein, a residue that is invariant cannot contain any information about the nature of a second residue.

Identification of significantly covariant residue pairs

Figure 3 shows the distribution of mutual information or covariance values for all 1,770 ($60 \times 59/2$) residue pairs in each of the three sequence sets. The skewed nature of the distribution, as noted in the color matrices, is obvious in these histograms. Most of the remaining analysis consists of examining the most highly covariant residue pairs, and the arrows on the three panels of Figure 3 show the range of mutual information values used to define the highly covariant pairs from each sequence set. Ten

Fig. 2. False color representation of mutual information for all residue pairs. The residue numbers are shown along the top and left axes. Each panel represents a different sequence set. **A:** 263 sequences. **B:** 173 sequences. **C:** 87 sequences. Mutual information values for each sequence set have been normalized to the extreme values of the color bar; red is the maximum amount of mutual information, purple is the minimum.

pairs were selected from the first sequence set, seven from the second, and four from the third. The cutoffs for these selections are arbitrary, but were chosen to be at or close to breakpoints in the distribution and to be approximately four standard deviations from the mean value of the distributions.

Figure 4 provides evidence for the statistical significance of the highly covariant residues pairs. One thousand sets of test sequences were generated, each set containing 263 sequences. Selection of each amino acid in the test sequences was performed randomly; the probability of picking a particular amino acid was

Fig. 3. Histograms showing the number of residue pairs having a given mutual information value. The relevant sequence set for each histogram is given in the boxes. Arrows indicate the range of values used to define the highly covariant pairs that are listed in Table 1.

determined by the frequency at which that amino acid is found at that residue position in the set of real sequences. For each residue pair, the mutual information value was calculated for each of the 1,000 randomized sequence sets. Figure 4 shows (1) the distribution of mutual information values for the real residue pairs, (2) the distribution of the mean values for residue pair mutual information derived from the randomized test sets, and (3) the distribution of the maximum values found for each of the residue pairs among the randomized test sets. The number of residue pairs having a given mutual information value (or

Fig. 4. Histogram demonstrating the nonrandom nature of covariant residues. The randomization test is described in the text.

mean and maximum value in 1,000 randomized sequence sets) are shown on a log scale to make more noticeable the number of real residue pairs that exceed the largest maximum value found for any of the 1,770 residue pairs in 1,000 randomized trials; the data for the real sequences is the same as that shown on a linear scale in Figure 3.

The residue pairs identified by the cutoffs shown in Figure 3 are listed in Table 1. The table lists 16 residue pairs because some of the 21 residue pairs identified as being highly covariant were on the lists derived from more than one sequence set. The numbers in each column of the table show the percentile rank for each residue pair in each of the three sequence sets. The effect of eliminating sequences that contribute significantly to the bias of the previous set is evident from this table. The possible structural or functional significance of several of these highly correlated pairs is considered below.

Salt-bridges identified by covariance

Residues 19 and 30 are the most strongly correlated pair of residues in the largest set of homeodomain sequences. Table 2 shows that most of the amino acid pairs found at 19-30 have oppositely charged side chains, and Figure 5 shows that, in the engrailed homeodomain, E19 and R30 form a salt-bridge (Clarke et al., 1994). One other pair in the list of 16 highly correlated residue pairs was also found to be dominated by an amino acid pair of opposite charge, and Figure 6 shows that amino acids

Fig. 5. Stereo drawing of the engrailed structure with residues E19 (light) and R30 (dark) indicated (Clarke et al., 1994). This and other structures in this paper were drawn using MOLSCRIPT (Kraulis 1991).

Table 1. Well correlated residue pairs^a

Residue pair		Number of sequences		
		263	173	87
19	30	1 ^{b,c}	3	15
28	54	1 ^{b,c}	3	4
44	46	1 ^{b,c}	1 ^{b,c}	1
8	43	1 ^b	3	4
46	47	1 ^b	1 ^b	6
26	46	1 ^b	1	37
26	47	1 ^b	1 ^b	24
44	47	1 ^b	1 ^b	10
46	50	1 ^b	1 ^{b,c}	3
26	44	1 ^b	1	32
12	50	1	1 ^{b,c}	2
17	52	2	1 ^b	1
31	42	3	1	1 ^b
6	47	47	1	1 ^b
8	25	8	1	1 ^b
46	54	3	5	1 ^b

^a The percentile rank for each of the 16 highly covariant residue pairs is given for each sequence set. For example, residue pair 19-30 is the single most covariant pair in the complete set of 263 sequences, thus it is within the top 1% of the 1,770 possible residue pairs. It is in the top 3% of pairs in the sequence set of 173 sequences and the top 15% of pairs in the sequence set of 87 sequences. The ordering of residue pairs was done as follows. The 10 residue pairs selected in the set of 263 sequences were ordered on the basis of mutual information value. Those pairs selected in the second set of sequences that were not selected in the first set (12-50 and 17-52) were appended to this list in order of mutual information value. The final four residue pairs are the four selected in the set of 87 sequences, and are appended to the list in the rank order of their mutual information value.

^b Indicates that the residue pair was within the cutoff value for inclusion in the list as defined in Figure 3 for each of the three sequence sets.

^c The three residue pairs having the greatest mutual information in that sequence set. Sequences containing the amino acid pair contributing most to each of these three residue pairs were eliminated in the next-smallest sequence set.

at these positions (17-52) can also form a salt-bridge (Qian et al., 1994). Interestingly, though, many homeodomain sequences contain lysines at both 17 and 52 rather than amino acids of opposite charge (Table 3). The significance of this is not clear.

Table 2. Amino acid pairs at residue pair 19-30

	Residue 30				Total
	R	K	E	Other	
Residue 19					
E	120	15	1	3	139
K	0	0	16	1	17
H	1	0	15	3	19
R	0	0	13	1	14
Other	4	0	25	45	74
Total	125	15	70	53	263

Fig. 6. Stereo drawing of ftz with residues E17 (light) and R52 (dark) (Qian et al., 1994). For this drawing, several residues were deleted for clarity from the ends of the structure. The view is down the third helix with the C-terminal end toward the reader.

Mutually exclusive phosphate-contact residues

Residues 31 and 42 are significantly more covariant than any other residue pair in the set of 87 sequences (Fig. 2; Table 1). This pair is also within the top 1% of pairs in the set of 173 sequences. Table 4 shows that the salient feature of the amino acid pairs at 31-42 is that most sequences have a positively charged amino acid at residue 31 or at residue 42 but not both. A rationalization for this correlation is shown in Figure 7, which shows the protein-DNA complexes of engrailed and of MAT α 2. R31 of engrailed and R42 of MAT α 2 each make contact with the same phosphate, suggesting there may be a functional constraint against having positively charged amino acids at both residues.

DNA-binding and the linkage between structural residues and functional residues

Among the 16 highly covariant pairs, 8 are part of a single network of covariance. Figure 8A summarizes the set of connections among these residues, and Figure 8B shows where the residues occur in the structure of the engrailed homeodomain bound to DNA. With the exception of residue 26, each of the residues is on the outside surface of the third helix, which lies in the major groove of DNA. Each of these residue positions has been implicated in DNA binding by homeodomains, either from crystallographic studies, mutagenesis studies, or both.

Residue 26 is not in contact with DNA. It lies in the loop between the first and second helices, and faces toward the back surface of helix 3. In the engrailed-DNA complex, residue 26 is a leucine that makes contact (<4.0 Å) with Ile 45 and Phe 49 on the interior surface of the third helix. The inclusion of this residue in a set of covariant residues known to be involved in DNA binding suggests that subtle structural changes accompa-

Table 3. Amino acid pairs at residue pair 17-52

	Residue 52			
	R	K	Other	Total
Residue 17				
E	117	0	6	123
K	3	12	5	20
Other	12	5	13	30
Total	132	17	24	173

Table 4. Amino acid pairs at residue pair 31-42

	Residue 42			Total
	R	K	Other	
Residue 31				
R	1	0	40	41
K	0	0	19	19
Other	1	19	7	27
Total	2	19	66	87

nying substitutions at residue 26 may have an effect on the DNA binding properties of residues in helix 3. Based on the set of pairwise correlations among residues 26, 44, 46, 47, 50, and 54, a small number of consensus sequences can be obtained for these six residues, and most homeodomain sequences closely match one of these consensus sequences (Table 5). Most sequences contain either a branched aliphatic amino acid or proline at residue 26, and there is a tendency to have rather different residues

in the DNA-recognition helix if the residue is proline rather than leucine, isoleucine, or valine.

Highly covariant residues without obvious structural or functional rationales

The remaining covariant pairs (28-54, 8-43, 12-50, 8-25, and 6-47) are not obviously related to one another in structural or functional terms. These pairs may be examples of the artifactual covariance that is possible as a result of variable selective pressures and the biased sampling of sequences. Alternatively, the physical or functional explanation for covariance could simply be too subtle.

Examination of covariance in residue sets that are expected to show functional linkage: N-terminal arm residues

Instead of looking at the most covariant positions and asking whether covariance can be understood in structural or functional terms, we can also look specifically at residues that are in physical proximity or seem likely to share a functional role and ask

Fig. 7. Stereo drawings of an engrailed-DNA complex (top, (Kissinger et al., 1990)), and of a MAT α 2 complex (bottom, (Wolberger et al., 1991)) R31 of engrailed contacts the analogous phosphate to that contacted by R42 of MAT α 2.

Fig. 8. **A:** Schematic showing the network of residues that are highly covariant. All except residue 26 lie on the DNA-exposed surface of helix 3. **B:** Stereo drawing of an engrailed-DNA complex (Kissinger et al., 1990). The view is down helix 3, with the C-terminal end pointing into the page. The side chains lying in the major groove on the surface of helix 3 are those of residues 44, 46, 47, 40, and 54. Residue 26 is in the loop at the top of helices 1 and 2.

to what extent they covary. One set of residues that might be expected to show interesting patterns are those in the flexible N-terminal arm of the homeodomain, which makes contacts in the minor groove of the DNA. Inspection of the mutual information values in Figure 2 shows that residues 1-7 are generally not strongly correlated with one another compared with other residue pairs. An exception is residue pair 5-6, which is the sixth most covariant pair in the set of 87 sequences (yellow square in upper left corner of Fig. 2C). Pair 5-6 is also the most covariant pair among arm residues in the set of 173 sequences. The other pairs of arm residues within the top 10% of pairs for this sequence set are 3-6, 3-7, and 6-7, although these are not as obvious as 5-6. Residue pairs 3-7 and 6-7 are also among the top 10% of all pairs in the full set of sequences and are the only such arm residue pairs in that set. These correlations among residues 3, 6, and 7 are of interest because replacement of residues 3, 6, and 7 of the Ubx homeodomain with those of Abd-B has been shown to alter the DNA binding specificity from that of Ubx to something more like Abd-B (Ekker et al., 1994).

Table 5. Sequence subfamilies defined by residue 26 and correlated helix 3 residues^a

26	No. sequences	44	46	47	50	54	No. sequences matching 4/5
L, I, V	194	Q	K	I	Q	M	167
P	67	V	R	V	C	Q	17
		R	Q	V	S, A, K	A	17

^a A sequence is considered to belong to the subfamily if four of five of the correlated helix 3 residues match the pattern in the table. For example, 194 of the 263 sequences have L, I, or V at residue 26 and, of these, 167 have at least four of the residues Q44, K46, I47, Q50, and M54. Two subfamilies of helix 3 residues correlate with P26.

Examination of covariance in residue sets that are expected to show structural linkage: Hydrophobic core residues

A second class of amino acids that might be expected to show some degree of covariance are the hydrophobic core residues. Some analyses of sequence families show a weak correlation between the degree of covariance and the proximity of two residues in the core (Gobel et al., 1994; Shindyalov et al., 1994). Figure 9 shows the relationship between covariance as defined in this paper and $C\beta$ - $C\beta$ distances for hydrophobic core residues in the homeodomain family. Although it looks as if there

Fig. 9. Scatter plot of the relationship between mutual information and $C\beta$ - $C\beta$ distances for all pairs of buried residues in the engrailed structure. Residue 48 was excluded from the analysis because it is invariably tryptophan. The residues included in this analysis were identified on the basis of solvent accessibility and are residues 8, 16, 20, 26, 34, 38, 40, 45, and 49 (Clarke et al., 1994). The correlation coefficient is 0.43 for an inverse relationship between distance and covariation. Negative values are not possible using the standard definition of mutual information (Korber et al., 1993), but are possible using the weighted form adopted here.

might actually be a slight negative correlation between proximity and covariation, the correlation is not significant.

Discussion

Real sequence sets are likely to be unrepresentative of the much larger set of sequences that could, in principle, meet the selection criteria that created the known sequence set. Problems arising from biased sampling and from differing selective pressures were discussed in the introduction of this article. Whether these artifactual effects can be distinguished from true evolutionary covariation will determine, in part, the extent to which covariation can be used as a tool in structure prediction or as an indicator of functional significance.

In this study, no correlation was found between residue covariation and residue proximity in the hydrophobic core. However, the possibility of using covariation as an indicator of structural proximity has been tested recently on a number of other protein sequence families, and weak, but possibly useful, correlations have been found in some cases (Gobel et al., 1994; Shindyalov et al., 1994). There is some disagreement as to whether such pairwise correlations provide additional information beyond that of a simple conservation metric, but differences in methodology make comparisons difficult (Taylor & Hatrick, 1994).

Part of the weak correlation between physical proximity and covariation can undoubtedly be ascribed to noise from the non-ideality of the sequence set. However, it is also possible that the connection between physical proximity and structural "linkage" is itself quite weak. Because it is the structural and functional linkages between two residues that determines how well correlated they would be in an ideal sequence set, a weak correlation between proximity and structural linkage would put an upper limit on our expectations for the correlation between proximity and covariation. Salt-bridges would be an exception to this idea because it is clear what the nature of the structural linkage is in this case, and the linkage has a direct relationship to residue proximity.

The identification here of two highly correlated residue pairs that can form salt bridges suggests that strongly correlated pairs showing covariation patterns consistent with salt-bridges might be a useful structural constraint in building models of proteins. Residue 19 is near the C-terminal end of helix 1 and residue 30 is near the N-terminal end of helix 2. Residues 17 and 52 are in helices 1 and 3, respectively. These two structural constraints, although nowhere near the number required to define the structure, would presumably eliminate a large number of alternative structures that might otherwise be proposed if the homeodomain structure were not known. Salt-bridges are likely to be relatively weak constraints, however, because of the length and flexibility of the side chains involved.

Instead of trying to use covariant residues to predict structure, covariance can be used to provide evidence for the functional significance of known structural features. For example, E19 and R30 make contact in the engrailed homeodomain, but the significance of this salt-bridge could not normally be assessed in the absence of mutagenesis and denaturation experiments. Covariance demonstrates that residue positions 19 and 30 influence the selective advantage of each other, providing strong evidence for the energetic significance of the observed salt-bridge.

The significance of a covariant residue pair is clearly greater if particular amino acid pairs have arisen and been selected for multiple times in evolution (Korber et al., 1993). Such seems to be the case, for example, with the 19-30 residue pair. However, covariance can also exist if there are combinations of amino acids that appear to have arisen once in evolution, but that are unusually conserved within the subfamily of sequences descended from that event. Indeed, most of the highly covariant residue pairs examined here fit this description to some extent. This does not mean, however, that covariation of this type is necessarily artifactual and without a structural or functional basis. The pairs examined here are unusually well correlated, and to say simply that covariation is an artifact of conservation within a subfamily begs the question of why the amino acids at these positions are both distinct from those in other sequence subfamilies and unusually highly conserved within the subfamily.

In the absence of evidence for multiple evolutionary origins, it is an open question whether a covariant pair is artifactual or has a real structural or functional origin. In some cases, the structural rationale is so compelling that real covariation seems likely. One example of such a rationale is the observation that an arginine at residue 31 in engrailed contacts the same phosphate as an arginine at residue 42 of MAT α 2. It seems very likely that the apparent selection against having a basic amino acid at both positions in the same sequence is related to this structural observation. On the other hand, residue pairs such as 12-50 have no obvious basis in structure or function, and it is quite possible that such pairs are the consequence of separate selective pressures on each residue. Perhaps improved methods for identifying truly significant correlations will be developed as more sequence families become large enough for studies of this type.

Acknowledgments

This work was supported by the Markey Center for Macromolecular Structure and Function at Johns Hopkins. I thank Steve Soisson for making corrections to the list of sequences.

References

- Billeter M, Qian Y, Otting G, Muller M, Gehring W, Wuthrich K. 1993. Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. *J Mol Biol* 234:1084-1097.
- Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. 1994. Structural studies of the engrailed homeodomain. *Protein Sci* 3:1779-1787.
- Egger SC, Jackson DG, Von Kessler DP, Sun BI, Young KE, Beachy PA. 1994. The degree of variation in DNA sequence recognition among four *Drosophila* homeotic proteins. *EMBO J* 13:3551-3560.
- Gobel U, Sander C, Schneider R, Valencia A. 1994. Correlated mutations and residue contacts in proteins. *Proteins Struct Funct Genet* 18:309-17.
- Kappen C, Schughart K, Ruddle FH. 1993. Early evolutionary origin of major homeodomain sequence classes. *Genomics* 18:54-70.
- Kissinger C, Liu B, Martin-Blanco E, Kornberg T, Pabo C. 1990. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell* 63:579-590.
- Klemm J, Rould M, Aurora R, Herr W, Pabo C. 1994. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA binding modules. *Cell* 77:21-32.
- Korber BT, Farber RM, Wolpert DH, Lapedes AS. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc Natl Acad Sci USA* 90:7176-7180.
- Kraulis P. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr* 24:946-950.

- Laughon A. 1991. DNA binding specificity of homeodomains. *Biochemistry* 30:11358-11367.
- Neher E. 1994. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA* 91:98-102.
- Qian Y, Billeter M, Otting G, Muller M, Gehring W, Wuthrich K. 1989. The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: Comparison with prokaryotic repressors. *Cell*:573-580.
- Qian YQ, Furukubo-Tokunaga K, Resendez-Perez D, Muller M, Gehring WJ, Wuthrich K. 1994. Nuclear magnetic resonance solution structure of the fushi tarazu homeodomain from *Drosophila* and comparison with the Antennapedia homeodomain. *J Mol Biol* 238:333-345.
- Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7:349-358.
- Taylor WR, Hatrick K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng* 7:341-348.
- Wolberger C, Vershon A, Liu B, Johnson A, Pabo C. 1991. Crystal structure of a MAT α 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* 67:517-528.