Protein Engineering vol.15 no.11 pp.871–879, 2003

An analysis of protein domain linkers: their classification and role in protein folding

Richard A.George[1] and Jaap Heringa[2]

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill NW7 1AA, UK. [1]Informatica Ltd, 60 Charlotte Street, London W1T 2NV, UK

[2]To whom correspondence should be addressed at Bioinformatics, Dept. Computer Science, Faculty of Science, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
E-mail: jhering@nimr.mrc.ac.uk

Recent advances in protein engineering have come from creating multi-functional chimeric proteins containing modules from various proteins. These modules are typically joined via an oligopeptide linker, the correct design of which is crucial for the desired function of the chimeric protein. Here we analyse the properties of naturally occurring inter-domain linkers with the aim to design linkers for domain fusion. Two main types of linker were identified; helical and non-helical. Helical linkers are thought to act as rigid spacers separating two domains. Non-helical linkers are rich in prolines, which also leads to structural rigidity and isolation of the linker from the attached domains. This means that both linker types are likely to act as a scaffold to prevent unfavourable interactions between folding domains. Based on these results we have constructed a linker database intended for the rational design of linkers for domain fusion, which can be accessed via the Internet at http://mathbio.nimr.mrc.ac.uk.
Keywords: domain/linker/module/peptide/proline

Introduction

Many cellular processes involve proteins with multiple domains. The modular nature of proteins has many advantages, providing increased stability and new cooperative functions. Other advantages include the protection of intermediates within inter-domain clefts that may otherwise be unstable in aqueous environments and the fixed stoichiometric ratio of enzymatic activity necessary for a sequential set of reactions (Ostermeier and Benkovic, 2000). It is not surprising then that advances in protein engineering have come from creating multi-functional chimeric proteins containing modules from various proteins (e.g. Nixon *et al*., 1997).

Recent studies have shown that domain linkers can play an essential role in maintaining cooperative inter-domain interactions (for a review see Gokhale and Khosla, 2000). An example is the intramolecular interaction between the Src homology domains (SH2 and SH3) and the catalytic domains of Src family kinases, which results in repression of catalytic activity. Repression by the regulatory domain is nullified upon mutation of Trp260 to Ala within the linker separating the SH2 and kinase domain, which proves that the linker plays a crucial role in the coupling of the regulatory domains to the catalytic domain (LaFevre-Bernt *et al*., 1998; Briggs and Smithgall, 1999). Another example is polyketide synthase, for which it was shown that the 'assembly line' catalysis observed in consecutive domains is reliant on having appropriate linkers between the domains of this protein (Gokhale *et al*., 1999). In yet another example, Ikebe *et al*. (Ikebe *et al*., 1998) demonstrated that deletion of four amino acids from the linker connecting two sub-domains of phosphorylated smooth-muscle myosin leads to the termination of its actin translocating activity: Not only is the composition of this linker important, but so is its length. In general, altering the length of linkers connecting domains has been shown to affect protein stability, folding rates and domain–domain orientation (van Leeuwen *et al*., 1997; Robinson and Sauer, 1998).

Some previous studies have identified particular types of linker. For example, Q-linkers occur at the boundaries of functionally distinct domains in a variety of bacterial regulatory and sensory transduction proteins, including the nitrogen regulatory proteins NtrB, NtrC, NifA and NifL (Wootton and Drummond, 1989). Q-linkers are between 15 and 20 residues in length and are not strongly conserved in sequence in otherwise homologous proteins. They have a preference for Gln, Arg, Glu, Ser and Pro residues and adopt a coil structure. Insertion of amino acids within the Q-linker sequence of NtrC and NifA was found to have no effect on their function (Wootton and Drummond, 1989). However, when NtrC was expressed as two separate polypeptides consisting of the domains normally joined by the Q-linker, the construct failed to function. In this case the linker anchors the domains together.

Many studies of linker peptides in various protein families have come to the conclusion that linkers lack regular secondary structure, they display varying degrees of flexibility to match their particular biological purpose and are rich in Ala, Pro and charged residues (Packman and Perham, 1987; Radford *et al*., 1989; Argos, 1990; Perham, 1991; Russell and Guest, 1991; Robinson and Sauer, 1998; Dieckmann *et al*., 1999). Argos (Argos, 1990) carried out a statistical study of natural linkers with the aim to design independent linkers for gene fusion that would have a low likelihood of disrupting the folding of the flanking domains. He constructed a set of 51 linkers from visual inspection of 32 proteins. The amino acids Thr, Ser, Pro and Asp were found to be desirable linker constituents. The author concluded that the preferred linker amino acids are mostly hydrophilic, often polar and usually small. The majority, 59%, of the linker residues were in coil or bend structures with a mean length of 6.5 residues, but an average flexibility when compared to other protein regions. It was suggested that pentapeptides consisting of only Gly, Ser and Thr would make the best linkers for gene fusion; as these residues were most strongly preferred within natural linkers. Differing structures pointed to the importance of the amino acid order to achieve an extended and conformationally stable oligopeptide (Argos, 1990).

The analysis by Argos is now slightly outdated since the protein data set used was small and linker delineation had been performed manually. In addition to a much larger data set, we have developed an automated method to extract inter-

domain linkers from a data set of proteins of known 3D structure. We have analysed the amino acids' propensities in linkers and examined the preferred order of residues within linkers. We have also devised a linker database, which can be used as a starting point to engineer domain fusion.

Methods

Protein data set

We used the non-redundant protein set available from the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Structure/VAST) as our primary database (Matsuo and Bryant, 1999). This set is derived from an all-against-all BLAST comparison of all proteins of known structure. The proteins in this set have been grouped into similar sequences using single linkage clustering based on BLAST *P*-values $<10^{-7}$. Chains within a group are ranked according to completeness and resolution of their structures, leading to 2101 protein representatives (Matsuo and Bryant, 1999). Proteins containing membrane-spanning regions were manually removed from the NCBI set.

Domain assignment in the final set of 1867 proteins was achieved using the automatic method of Taylor (Taylor, 1999) which determines domains by their compactness in space. The method first assigns to each residue in the protein structure a numeric label. Initially, the sequence residue numbers are taken as labels, which are then iteratively changed based on the respective residue neighbourhoods. If a residue is surrounded by neighbours (within a given radius *r*) with, on average, a higher label than the central residue, its label is increased by one; otherwise it is decreased by one. This test and reassignment is made repeatedly to each residue in the protein, which results in compact regions evolving towards the same residue number. Upon convergence, the final domain boundaries are then assigned between such compact regions. From the obtained domain boundaries the associated linker regions were delineated as discussed in the next section.

Determining the linker region

For each protein, linker regions were determined by branching out from the domain boundaries assigned by the Taylor algorithm. Linker assignment ended when the branches became buried within the core of a domain or when a branch accumulated a length of 40 residues. Therefore, a maximum of 80 residues was allowed for any one linker, although none of the linkers in our set reached this limit.

The structural environments (core or exposed) for the amino acids in each of the isolated domains were determined by calculating residue solvent accessibility using DSSP (Kabsch and Sander, 1983). All solvent accessibilities were normalized using the estimated maximum values derived by Chothia (Chothia, 1976). Residues were classified as being within the core of a domain if they had a normalized solvent accessibility below 20%, otherwise they were deemed 'exposed' and classed as surface residues. Any linker that grew to within 15 residues of another linker or 15 residues from the N- or C-termini of its respective protein was discarded.

Assigning linkers to sets

The linkers were arbitrarily divided into several sets based on their length; small (less than six residues), medium (between six and 14 residues) and large (greater than 14 residues). The linkers where also classified by the number of linkers involved in connecting two domains; 1-linker (i.e. one linker connecting continuous domains), 2-linker, 3-linker and >3-linker sets. Analysis of linker conformation indicated that the linkers would be best split into two general types: helical and non-helical. Linkers with over 33% of their residues in helical structures as annotated by DSSP (Kabsch and Sander, 1983) are classified as helical, otherwise as non-helical. For compositional comparison of the linkers, further sequence sets were generated based on secondary structural units within the NCBI proteins. These sequence sets were divided into helical, strand and loops connecting secondary structure as defined by DSSP.

Calculating amino acid propensity

Linker propensities of individual residues and residue pairs were calculated for the extracted linkers. The single amino acid propensities were determined from the ratio of their occurrence in the linker set compared to its occurrence in the full protein set:

$$Pa = \frac{Nr_{i,l}/\Sigma_i Nr_{i,l}}{Nr_{i,t}/\Sigma_i Nr_{i,t}} \quad (1)$$

where *Pa* is the propensity for amino acid *i*, $Nr_{i,l}$ and $Nr_{i,t}$ are the number of amino acid type *i* in the linker set (*l*) and in the full protein set (*t*), respectively. $\Sigma Nr_{i,l}$ and $\Sigma Nr_{i,t}$ are the total number of amino acids in the linker set and in the full protein set, respectively. Amino acid pair propensities were calculated following the weighted method described by Crasto and Feng (Crasto and Feng, 2001) in their analysis of loops connecting secondary structure:

$$Pab = \frac{Pa + Pb}{2} \times \frac{[N_{Pab,l}]/[\Sigma_i N_{Pai,l} + \Sigma_i N_{Pib,l}]}{[N_{Pab,t}]/[\Sigma_i N_{Pai,t} + \Sigma_i N_{Pib,l}]} \quad (2)$$

where *Pab* is the residue pair propensity to be found for a given dipeptide *ab*. *Pa* and *Pb* are the individual propensities as calculated in equation (1) for residues *a* and *b*, respectively. $N_{Pab,l}$ and $N_{Pab,t}$ are the number of occurrences of the residue pair (*ab*) in linkers (*l*) and in the full protein set (*t*), respectively. $\Sigma_i N_{Pai,l}$ and $\Sigma_i N_{Pib,l}$ represent the occurrence of residue pairs *ai* and *ib* in the linkers, respectively, while $\Sigma_i N_{Pai,t}$ and $\Sigma_i N_{Pib,t}$ represent the number of residue pairs *ai* and *ib* in the full protein set, respectively. Linker preferences were calculated for all (20×20) residue pairs.

Results and discussion

From a data set of 638 multidomain protein chains, 1280 linkers were extracted, totalling 12 776 residues. Linkers are found to have an average length of 10.0 ± 5.8 residues (Figure 1). The linkers were split into several sets as described in Methods; small, medium and large linkers have an average length of 4.5 ± 0.7, 9.1 ± 2.4 and 21.0 ± 7.6, respectively. The non-helical and helical linker sets both have length distributions similar to the full set. The number of residues differs between linker groups with varying connectivity; the lengths for the 1-linker, 2-linker, 3-linker and >3-linker sets (based on the number of linkers involved in connecting two domains) decrease in their average number of residues as the number of connections increases, 11.4 ± 7.0, 9.5 ± 4.5, 8.7 ± 4.1 and 8.2 ± 3.9, respectively. This suggests that in proteins with discontinuous domains the relative domain movement is not only restricted as a result of multiple linkers but also from tighter links.

Linker residues are shown to be partially buried with an average normalized solvent accessibility of 26.7% compared

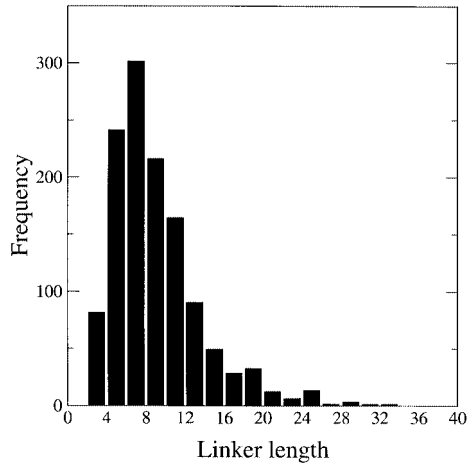

Fig. 1. Linker length distribution.

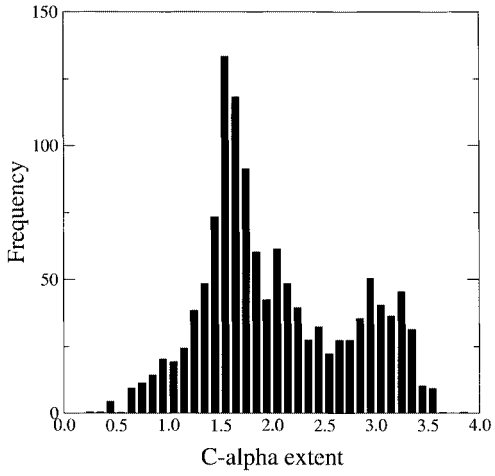

Fig. 2. Linker C_α-extent (residual translation) calculated by dividing the distance between two terminal C_α atoms with the number of residues minus one.

to 45.6% when the flanking domains are separated. Therefore, the relative burial of linkers after domain formation is 41.5%. Small linkers have a low average normalized solvent accessibility of 14.4% compared to 26.3% for medium sized linkers and 42.5% for large linkers. The larger the linker the more exposed it will be. The linker sets based on the number of linkers involved in connecting two domains, 1-linker, 2-linker, 3-linker and >3-linker, decrease in their average normalized solvent accessibility as the number of connections increases, 21.1, 19.0,16.8 and 13.8%, respectively.

Linker conformation

The C_α extent (or residual translation) of a linker is calculated by dividing the distance between two terminal C_α atoms with the number of residues minus one (Figure 2). The average C_α distance is 2.04 ± 0.76 Å per residue. Figure 2 shows that the distribution can be split into two, based on peaks at 1.5 Å and another at 3 Å, corresponding to helical and fully extended conformations, respectively.

Residues in helices are subject to steric constraints; their torsion angles are restricted to a smaller region of the Ramachandran plot. A helical linker might well be important in correct domain folding, acting as a rigid spacer to separate two domains. Some helical conformations may rapidly form during folding (Aurora *et al*., 1997), allowing the domains to fold independently without forming non-native interactions with the linker. Although the conformation of extended linkers would not appear to promote rigidity, there are compositional effects that correspond to this (see later).

Overall, the largest proportion of linker residues, 38.3%, adopt the α-helical secondary structure, 13.6% are in β-strands, 8.4% are in turns and the rest, 37.6%, are in coil or bend secondary structures. The helical proportion we find is markedly different to the observations made by Argos, where only 13% of linker residues adopted a helical conformation (Argos, 1990).

The residues in small linkers mainly comprise β-strand and coil secondary structures, 33.6 and 36.9%, respectively, only 21.0% are helical and the remainder, 8.5%, are in turns. Medium sized linkers have more residues in helical and coil secondary structures, 43.1 and 34.2%, respectively, while only 13.0% are found in β-strands, the remainder, 9.7%, are in turns. Large linkers are mainly in helical and coil secondary structures, 31.4 and 45.4%, respectively, while only 10.4% are found in β-strands, the remainder, 12.8%, are in turns. Short linkers are most distinct due to their relatively high β-strand and low α-helical content. Secondary structure type is not seen to deviate between the sets with varying number of linkers between two domains.

Hydrophobicity and domain separation

We analysed the average residue hydrophobicity for the linkers using Eisenberg's normalized consensus residue hydrophobicity scale, which ranges from 0 (hydrophilic) to 1 (hydrophobic) (Eisenberg, 1984). Overall, average hydrophobicity for linkers is 0.65 ± 0.09. Small linkers show an average hydrophobicity of 0.69 ± 0.11, while large linkers are more hydrophilic with 0.62 ± 0.08. The more exposed the linker, the more likely it is to contain hydrophilic residues. Longer linkers are independent and may allow larger domain motions. Both helical and non-helical linkers have average hydrophobicities similar to the full set.

Hydrophobicity does not deviate much from the mean when considering the number of linkers connecting domains, although >2-linkers are slightly more hydrophobic. However, the average number of residues differs between linker groups with varying connectivity; as the number of connections increases the average number of residues decrease (see above). Also, the average distance measured in angstroms (Å) between linker N-termini and C-termini C_α residues decreases significantly, 17.6 ± 10.3, 15.6 ± 6.0, 14.3 ± 5.5 and 13.8 ± 4.9, respectively. More linker connections between two domains means greater hydrophobicity, less distance and tighter interaction between domains, such that domain–domain movement is likely to be severely restrained.

Amino acid propensity for linkers

The preferred linker amino acids observed in the majority of the linker sets are Pro, Arg, Phe, Thr, Glu and Gln, in order of decreasing preference (Table I). Previous studies highlighted Ala to be a desirable linker constituent (see Introduction), surprisingly our study shows Ala to be undesirable in linkers. Although Ala does have a high occurrence in our set of linkers, it also has the highest overall frequency of occurrence in proteins and shows no linker preference. Argos (Argos, 1990) found Thr, Ser, Pro, Asp, Gly, Lys, Gln, Asn and Ala (in order of decreasing preference) to be the preferred residues in linkers, but it should be noted that Ala was only slightly preferred in the Argos study. Pro is the most preferred of the linker residues

Table I. Linker propensities

	All	1-linker	2-linker	3-linker	Small	Medium	Long	Helical	Non-helical
Pro	1.299	1.362	1.266	1.332	1.241	1.314	1.309	0.8	1.816
Arg	1.143	1.129	1.137	1.069	1.131	1.132	1.154	1.239	1.038
Phe	1.119	1.122	1.11	0.981	1.368	1.121	1.058	1.09	1.151
Leu	1.085	1.11	1	1.193	1.192	1.106	0.994	1.276	0.885
Glu	1.051	1.054	1.139	0.992	0.736	1.053	1.115	1.199	0.9
Gln	1.047	1.092	0.916	1.111	0.861	0.999	1.2	1.124	0.968
Met	1.032	0.923	1.077	0.998	1.369	1.093	0.782	1.171	0.878
Thr	1.017	1.023	1.018	0.992	0.822	0.988	1.11	0.832	1.189
His	1.014	0.949	1.109	1.034	0.973	1.054	0.992	1.012	1.05
Tyr	1	0.902	1.157	1.12	0.836	1.09	0.866	1.075	0.945
Ala	0.964	0.974	0.938	1.042	1.065	0.99	0.892	1.092	0.843
Val	0.955	0.923	0.959	1.001	1.14	0.957	0.9	0.908	0.999
Ser	0.947	0.932	0.956	0.984	1.097	0.911	0.986	0.886	1.003
Asn	0.944	0.988	0.902	0.828	0.762	0.873	1.144	0.927	0.956
Lys	0.944	0.946	0.952	0.979	0.478	1.003	0.944	1.008	0.893
Ile	0.922	0.928	0.986	0.852	1.189	0.95	0.817	0.912	0.946
Asp	0.916	0.892	0.857	0.97	0.836	0.915	0.925	0.919	0.906
Trp	0.895	0.879	0.971	0.96	1.017	0.939	0.841	0.981	0.852
Gly	0.835	0.845	0.892	0.743	1.022	0.785	0.917	0.698	0.978
Cys	0.778	0.972	0.6856	0.5	1.015	0.644	1.035	0.662	0.896

Propensities >1.1 are highlighted.

in our data set. It is likely to be favoured because it has no amide hydrogen to donate in hydrogen bonding and therefore, structurally isolates the linker from the domains.

Table I shows the general linker amino acid preferences as well as those for the other classes. Medium sized linkers show preferences similar to those of the entire set. The long linkers have an increased propensity for Cys, Asn and Gln, and a decreased preference for the hydrophobic amino acid Met. The short linkers show increased propensities for hydrophobic residues and decreased propensities for polar and acidic residues. As for the general hydrophobic tendency, the linker groups according to the number of linkers connecting two domains do not show significant differences in amino acid propensities.

The amino acid propensities for helical and non-helical linker sets are shown in Table I. The non-helical set consists of 51% of all linkers and shows the highest propensity for Pro residues, 1.81. High preferences are also shown for Thr, Phe, His and Arg. Having many Pro residues will cause a high degree of stiffness in the linker, which could be a requirement for the correct folding of two domains, as in the case of helical linkers by their inherently stiffer conformation. This is confirmed by the observed composition in helical linkers, which show a preference for Leu, Arg, Asp, Met and Gln. Pro is avoided in the helical set, with a propensity of only 0.80. These values generally correspond to the α-helical propensities described by Chou and Fasman (Chou and Fasman, 1974), although Arg, Gln and Pro show an increased propensity in the helical linkers and Ala, Glu, Met and Val show a decrease in propensity. A disruption in the normal pattern of hydrogen bonding and conformational constraints on the helix will allow larger torsion angle changes and such disruptions can be caused by proline residues. Out of the linkers that are fully helical, 8.2% (15/184) contain a central proline, compared to under 2% for helices in general and therefore allowing the possibility for a hinge mechanism within these linkers.

Linkers versus loops and secondary structure

To compare our linker observations, a database of protein loops connecting secondary structures, as assigned using DSSP (Kabsch and Sander, 1983), was generated from the NCBI protein set. Using this database we found that the composition of inter-domain linkers is distinct from intra-domain loops connecting secondary structures with an amino acid propensity correlation coefficient of 0.07. Residues Pro, Gly, Asp, Asn, His, Ser and Thr (in order of preference) are preferred in loop regions. In contrast, Gly, Asp, Asn and Ser are the least preferred within linkers, while His and Thr have no preference. Proline shows a high preference in both linker and loop sets, but will play a different role in each. A proline residue within a loop is likely to be involved in a tight proline turn, whereas only few proline turns were observed in our linker set.

All intra-domain residues that adopt an α-helical and β-strand structure were also extracted from the NCBI protein set and used to compare the composition of the helical and non-helical linker sets. The helical linkers have an amino acid composition similar to the internal helices taken from the protein set, with a high correlation coefficient of 0.85. The non-helical linkers have a composition most like coil structures, but with a lower correlation coefficient of 0.54. In agreement with the deviating amino acid composition and the above observation for proline, non-helical linkers have had to adapt more to their linker role in order to attach the relative independence and stiffness inherent in their helical counterparts.

χ^2 comparison of linker sets

A χ^2 test was used to analyse the significance of trends of the amino acid linker composition between the various linker sets (Table II). The sets representing linkers by the number required to connect two domains have no significant compositional differences tested at the 0.1% significance level and therefore suggest that there are no additional amino acid requirements

Table II. χ^2 test between linker sets

	All	1-linker	2-linker	3-linker	Small	Medium	Long	Helical	Non-helical
All	x								
1-linker	8.2	x							
2-linker	10.1	19.6	x						
3-linker	11.2	16.6	15.6	x					
Small	37.1	37	35.1	35.8	x				
Medium	8.5	21.9	9.7	6.8	39.6	x			
Long	27.3	13.7	28.3	29.2	50.2	46.4	x		
Helical	94.1	89.5	63.4	34	58.2	70.7	98.4	x	
Non-helical	84.6	58.6	48.8	47.5	45.9	94	45.1	265.5	x
NCBI	136	81.8	42.3	37.3	41.4	122.8	58	182.4	251
Coil	1109.9	543.9	329	198.9	148.2	850.3	236.3	1003.9	1752.1
a-helical	1654.5	966.7	536.9	255.1	185.6	1068.9	657.7	274.4	1957.5
b-strand	2192.2	1286.5	665.9	389.3	159.3	1511.3	858.8	1056.3	393.1

Null hypothesis: two compared sets have similar amino acid composition. Shaded cells identify pairs of sets that are significantly different, measured at the 0.1% significance level.

that define the number of linkers connecting two domains. As the number of linkers connecting domains increases, the composition of the corresponding linkers becomes more like that found in the general non-redundant protein set, which implies that these linkers become more like inter-domain segments.

The composition of small linkers is not significantly different to that in the non-redundant NCBI protein set at the 0.1% significance level. Small and medium linkers have a similar composition, which differs from long linkers. However, both medium and long linkers have significantly different compositions compared to the NCBI protein set. Also, the helical and non-helical linker compositions are very different from each other and consequently from the linkers as a whole. Although the composition of helical linkers was found to correlate most with internal helices derived from the protein set, the χ^2 test shows a significant difference between them. These observations suggest that a successful method to predict linker location in a sequence, as a method of domain delineation, must take account of linker type. At the very least, predictions must be made separately for helical and non-helical groups. Also, it is likely that small linkers would be very difficult to identify given their overall similarity to residues in the non-redundant protein set.

Dipeptide propensities for linkers

We have tried to refine our observations regarding compositional effects by calculating dipeptide propensities for all linkers and medium sized linkers, as shown in Tables III–VI. The small and large linker sets do not have enough representatives to permit reliable propensities. In the sets provided, residue pairs are well represented, with a mean number of 28.5 ($\pm$18.4) for each possible pairing. However, Met–Cys pairings are not observed in the linkers, but Cys–Met shows the second highest preference in the medium linker set, although there are only four occurrences. Both Cys and Met amino acids are rare. Pro–Pro pairs are the most common in the full and medium linker sets. Again, this suggests linkers prefer to be isolated from the rest of the protein.

The most favoured pairs in non-helical linkers are Pro–Pro, Trp–Trp, Met–His, Gln–Pro and Pro–Leu (Table V). Those favoured in the helical group are Cys–Met, Arg–Met, Arg–Lys, Gln–Arg and His–Arg (Table VI). The non-helical preferred pairs, with preferences >1.3, are of average hydrophobicity (63% Eisenberg scale). The least preferred pairs in this set, with propensities < 0.7, are often hydrophobic (71% Eisenberg scale). The helical set prefers amino acid pairs (propensities >1.3) that are slightly more hydrophilic (59% Eisenberg scale), and disfavours amino acids (propensities <0.7) that are more hydrophobic (69% Eisenberg scale). Prolines are observed in the helical set when paired with His or Met, but in cases of His, only when His precedes Pro.

Asp–Pro is one of the most frequent Xaa–Pro pairings found in helical linkers, but has a low occurrence in the non-helical linker set. Conversely, Glu–Pro is avoided in the helical linkers, but has a very high preference in the non-helical linker set. Although Asp and Glu both have acidic side chains and are often seen to be inter-changeable through protein evolution, they show different propensities to be in an internal strand or helical conformation (Chou and Fasman, 1974), i.e. both show a dislike to be in strand but Glu has a strong preference for helix while Asp has a strong preference for coil.

In Crasto and Feng's analysis of loops (Crasto and Feng, 2001) residue pairs (a–b) were identified that were highly favourable for loop conformation, whereas their reversed complement (b–a) had little or no preference (referred to as asymmetric dyads). Using Crasto and Feng's method of selection, we have identified asymmetric dyads with a high propensity for linkers (Table VII). Residues His and Trp are involved in a large number of dyads. Interestingly, His–Pro, Lys–His, Lys–Tyr and Phe–Ser pairs in the helical set have their reverse complement in the non-helical set, suggesting that the order of these amino acids are of key importance in determining the overall structure of the oligopeptide.

The importance of proline

Proline is the most preferred amino acid type in both linker and loop regions. Proline is unique among protein residues as it is a cyclic imino acid with no amide hydrogen to donate in hydrogen bonding. Therefore, it cannot fit into the regular structure of either α-helix or β-sheet and is a common 'breaker' of secondary structure. It is the second most common residue in the first position of a helix, although it does infrequently occur in central positions (Chakrabarti and Chakrabarti, 1998). The proline ring pushes away the preceding turn in a helix by

Table III. Amino acid pair propensities over all linkers

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.81	0.68	0.79	1.2	1.01	0.85	1.11	0.76	0.87	0.81	1.13	0.96	1.35	1.31	1.18	0.86	0.92	0.91	0.95	0.98
C	0.77	0.22	0.51	0.79	0.77	0.53	1.42	0.44	0.7	0.43	1.34	0.72	0.5	0.68	0.9	0.3	0.81	0.38	0.53	0.72
D	0.72	0.73	0.68	0.88	1.12	0.56	1.35	0.9	1.05	0.94	0.83	1.18	1.19	1.03	1.55	0.96	0.88	0.86	1.1	1.02
E	1.15	0.8	0.88	1.11	1.19	0.9	1.02	1.12	1.01	1.16	1.08	0.84	1.39	1.25	1.46	1.05	1.11	1.18	0.79	0.76
F	0.8	0.8	1.05	1.26	1.63	1.09	0.59	1.27	0.91	0.89	0.69	1.51	1.24	1.47	0.91	1.14	0.9	0.88	1.47	1.37
G	0.95	0.76	0.68	0.68	1.11	0.62	0.74	0.68	0.72	0.86	1.35	0.98	1.01	1.11	0.99	0.84	0.97	0.75	0.8	0.9
H	0.88	0.37	0.74	0.97	1.06	0.86	1.04	0.75	1.39	1.19	1	1.16	1.27	0.93	1.66	0.94	1.26	0.99	0.96	1.33
I	0.76	1.07	0.93	1.02	0.76	0.72	0.82	1	0.79	0.86	1.13	1.04	1.27	0.94	0.99	0.82	0.81	0.87	0.6	0.73
K	0.86	0.93	0.9	1.11	0.82	0.85	1.22	0.73	0.81	1.12	1.17	0.89	1.41	1.22	1.33	0.91	1.04	0.85	0.64	1.25
L	1.08	0.62	1.2	1.28	1.38	0.88	1.18	0.92	1.12	0.97	0.98	1.11	1.47	0.9	1.04	1.01	0.84	0.85	1.01	0.91
M	1.39	0	1.08	0.64	1.15	0.85	1.56	0.89	1.11	0.92	0.69	0.52	1.76	0.75	1.1	0.82	0.95	0.97	0.83	1.06
N	1.04	0.71	0.98	0.93	1.1	0.53	1.01	1.06	0.93	1.17	0.4	0.75	1.23	1.34	1.06	0.79	0.85	1.05	1.04	1
P	1.44	0.71	1.17	1.25	1.22	1.19	1.49	1	0.92	1.6	1.35	1.2	1.91	1.22	1.44	1.37	1.36	1.4	1.35	1.31
Q	0.73	0.59	1.1	1.17	1.19	1.13	1.04	0.69	1.37	1.07	0.89	1.04	1.63	0.89	1.16	1.44	1.01	1.33	0.8	0.99
R	0.93	0.79	1.21	1.3	1.09	1.07	0.95	1.41	1.5	1.23	1.56	1.06	1.21	1.33	1.04	1.05	1.29	1.17	0.97	1.07
S	0.83	0.75	0.9	1.19	1.1	0.79	0.86	0.95	0.98	1.14	0.82	0.67	1.27	0.83	1.06	0.99	1.15	0.88	0.75	0.69
T	1.32	0.55	1.14	1.15	0.86	0.86	1.18	0.77	0.75	0.92	1.22	1.08	1.45	0.94	1.41	0.9	1.22	0.73	0.82	1.17
V	0.83	0.38	0.89	0.92	1.02	0.88	1.3	0.76	1.03	1.13	0.88	0.81	1.49	0.98	1.1	0.87	1	0.68	0.54	1.05
W	0.47	0.44	0.51	0.93	0.51	1.4	0.46	0.42	0.97	0.66	0.96	0.98	0.61	0.58	1.13	1.07	0.95	1.07	1.43	0.62
Y	0.94	0.7	0.95	1.22	0.95	1.02	0.94	0.99	1.16	1.18	0.81	0.59	1.29	1.47	1.23	0.9	1.13	0.82	0.97	0.59

Amino acids in the rows precede amino acids in the columns. Propensities >1.30 are highlighted. Sample size is 28.5 ± 18.4 per residue pair.

Table IV. Amino acid pair propensities for medium sized linkers

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.71	0.6	0.88	1.09	1	0.84	1.33	0.8	1.07	0.88	1.17	0.8	1.43	1.07	1.25	0.75	0.91	1.03	1.12	1.33
C	0.34	0.35	0.44	0.51	0.23	0.5	0.72	0.32	0.38	0.31	2.04	0.72	0.49	0.34	0.73	0.34	0.67	0.56	0.83	0.53
D	0.76	0.8	0.58	1.05	1.25	0.49	0.99	0.9	1.1	1.07	0.82	1.02	1.02	1.07	1.76	1.04	0.83	0.71	1.13	1.16
E	1.28	0.92	0.94	1.16	0.95	0.79	1.12	1.01	1.13	1.19	1.19	0.79	1.25	1.08	1.61	0.9	1.15	1.3	0.49	0.72
F	0.98	0.24	1.13	1.23	1.54	0.89	0.92	1.21	1.11	0.72	0.54	1.6	1.19	1.56	0.94	1.14	0.88	0.85	1.72	1.42
G	1.09	0.63	0.65	0.66	1.07	0.55	0.76	0.71	0.59	0.96	1.3	0.65	1.01	0.87	0.79	0.66	0.93	0.6	0.86	1.07
H	0.82	0.58	0.8	0.85	1.38	0.86	1.45	0.84	1.55	1.22	1.29	1.22	1.04	1.32	1.13	0.95	1.43	0.68	0.74	1.72
I	0.79	1.05	1.09	1.2	0.84	0.65	0.81	1.1	0.65	0.77	1.32	1.07	1.11	1.12	0.99	0.83	0.81	0.98	0.46	0.65
K	0.85	0.45	0.97	1.16	0.75	0.97	1.29	0.89	0.76	1.36	1.49	0.77	1.47	1.36	1.43	0.97	0.98	0.97	0.65	1.33
L	1.22	0.8	1.26	1.22	1.7	0.79	1.06	0.92	1.09	0.95	0.92	1.26	1.73	0.82	1.07	0.91	0.66	0.83	1.43	1.13
M	1.22	0	0.85	0.82	1.21	0.59	1.23	0.69	1.37	1.22	0.54	0.7	1.9	0.68	1.4	0.84	1.52	0.93	1.3	1.18
N	0.95	0.36	1.04	0.83	1.09	0.49	1.01	0.86	1	1.13	0.48	0.49	1.2	0.96	0.88	0.74	0.88	1.04	1.1	0.95
P	1.41	0.53	1.16	1.21	1.21	1.15	1.55	1.28	1.08	1.45	1.21	1.31	2.08	1.09	1.71	1.35	1.33	1.42	0.95	1.36
Q	0.81	0	1.16	1.19	1.32	0.88	1.33	0.68	1.21	0.95	0.88	0.92	1.61	0.98	0.8	1.44	0.88	1.44	0.83	0.96
R	0.93	0.73	1.18	1.47	0.78	0.92	1.06	1.39	1.53	1.18	1.84	0.97	1.33	1.45	1.01	1	1.26	1.16	0.83	1.28
S	0.76	0.67	0.67	1.13	0.91	0.76	0.68	1.09	1.08	1.1	0.71	0.65	1.19	0.84	1	1.05	1.02	0.86	1.03	0.82
T	1.47	0	1.06	1.32	0.93	0.91	1.34	0.72	0.74	0.99	1.67	0.78	1.35	0.67	1.34	0.83	1.2	0.55	0.72	1.14
V	0.75	0.23	0.9	0.78	1.02	0.83	1.2	0.96	1.3	1.24	1	0.74	1.67	0.73	1.21	0.68	0.96	0.64	0.68	0.99
W	0.43	0.69	0.63	1	0.53	1.36	0.72	0.43	1.34	0.46	0.99	0.79	0.48	0.73	1.06	1.19	0.67	1.34	1.48	0.63
Y	1.07	0.21	1.05	1.32	1.18	1.14	1.09	1.03	1.41	1.05	1.04	0.54	1.17	1.94	1.07	1.16	1.35	0.92	1.24	0.61

Amino acids in the rows precede amino acids in the columns. Propensities >1.30 are highlighted. Sample size is 18.4 ± 12.4 per residue pair.

~1 Å producing a bend of 26° along the helix axis and breaking the hydrogen bonds at position $i - 3$ and $i - 4$ (Barlow and Thornton, 1988). Proline will introduce some motion into a helix, that enables a number of different conformations at that region (Pastore *et al*., 1989). Also, for transmembrane helices it has been suggested that proline induces essential hinge bending, required in ion channel gating (Tieleman *et al*., 2001). NMR studies suggest that proline-rich sequences form relatively rigid extended structures and show 'elbow bending' dynamics (Radford *et al*., 1987). Two prolines in a row favour the poly-proline, or collagen, conformation, which is extended but cannot form a β-sheet. Short proline rich sequences are stiff, with non-interacting connections. As suggested before, this is the most likely reason why proline is the preferred linker constituent, particularly in non-helical linkers. It cannot hydrogen bond to any surrounding amino acids, avoiding ordered structure formation and contact with the neighbouring domains.

Prolines can adopt two conformations, α ($\phi = -61°$, $\psi = -35°$) or β ($\phi = -63°$, $\psi = 150°$). The conformation of the proline is influenced by the preceding residue, and when the preceding residue is hydrophobic, proline generally favours the β-conformation (MacArthur and Thornton, 1991). This tendency is confirmed in our linker database, as hydrophobic residues are the preferred amino acid types at linker positions preceding prolines, leading to the extended conformation.

Peptide bonds between proline and its preceding residue (Xaa–Pro) typically exist in equilibrium between *trans* ($\omega = 180°$) and *cis* ($\omega = 0°$) conformation, with respect to the two successive C_α positions. The geometry is such that a *cis* proline always forms a 180° turn in the polypeptide, which is known as a type VI or *cis*-Pro turn. Such a turn in a linker would be disadvantageous, bringing the flanking domains in close proximity and possibly causing conflicts during domain folding. The disadvantage of *cis*-Pro turns is aggravated by the fact that conversion between *cis* and *trans* is very slow, requiring

Table V. Amino acid pair propensities for non-helical linkers

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.75	0.75	0.76	0.93	1.14	0.91	1.21	0.54	0.58	0.51	0.78	0.79	2.17	1.13	0.87	0.73	1.14	0.78	0.33	0.57
C	0.51	0.46	0.85	0.78	0.64	0.66	1.29	0.46	0.74	0.33	0	0.75	0.96	0.73	1.58	0.62	0.89	0.39	1.13	1.15
D	0.75	0.59	0.55	0.72	1.45	0.81	1.6	0.96	0.97	0.73	1.03	1.2	1.25	0.57	1.49	0.94	0.92	0.89	0.37	1.14
E	0.67	1.38	1	0.84	1.06	0.97	0.5	1.16	0.8	0.79	0.65	0.72	2.23	0.78	1.16	1.06	0.7	1.37	0.66	0.47
F	0.62	1.34	1.41	1.24	1.58	1.47	0.6	0.99	1.01	0.8	0.23	1.58	1.51	1.32	1.06	0.97	1.25	0.59	1.49	1.06
G	1.02	0.97	0.75	0.82	1.32	0.81	0.81	0.78	0.82	0.81	1.46	1.21	1.37	1.01	1.12	1.02	1.29	0.9	1.17	0.9
H	0.86	0	0.74	0.58	1.58	1.22	0.69	1.09	2.04	1.24	0.97	1.54	1.15	1.32	1.46	1.05	0.94	0.74	0.96	1.12
I	0.79	1.22	0.94	0.67	0.54	0.85	0.89	1.12	0.79	0.57	1.61	1	1.96	0.68	0.81	0.91	0.87	1.16	0.59	0.82
K	0.83	1.16	0.85	0.85	0.83	1.1	1.05	0.44	0.86	0.87	1.06	0.77	1.89	1.03	1.06	0.88	1.32	0.81	0.21	1.09
L	0.53	0.15	1.17	0.93	0.99	0.9	0.87	1.01	1.04	0.68	1.08	0.96	1.74	0.57	0.65	1	0.95	0.61	0.58	0.88
M	1.13	0	0.91	0.46	1	0.76	2.7	0.68	1.02	0.32	0.68	0.51	1.91	0.42	0.43	0.99	0.81	0.87	0	0.61
N	0.9	0.74	0.68	0.52	1.22	0.61	0.92	1	0.62	1.3	0.19	0.71	1.81	1.55	1.31	0.96	1.26	0.88	0.7	1
P	2.32	1.13	1.49	1.56	1.54	1.67	2.22	1.59	1.3	2.35	1.42	1.69	3.48	2.05	2.07	1.58	1.92	1.99	1.81	1.49
Q	0.56	0.99	0.84	0.85	0.8	1.08	1.15	0.4	1.37	0.93	0.43	1.19	2.63	0.76	0.24	1.58	1.02	1.52	0.54	1.12
R	0.85	1.06	0.72	1.2	0.87	1	0.84	1.54	0.78	0.96	0.77	1.19	1.83	1.03	0.83	1.29	1.62	1.22	0.88	0.75
S	0.8	0.61	0.82	1.05	1.38	0.95	1.09	1.13	1.03	0.99	0.97	0.7	1.81	0.86	1.23	1.05	1.4	0.88	0.54	0.68
T	1.21	0.95	1.39	1.23	1.04	1.08	1.53	1.05	0.85	0.77	1.14	1.44	2.09	1.27	1.72	1.01	1.59	0.84	0.87	1.33
V	0.58	0.47	0.93	0.87	1.23	1.12	1.52	0.78	0.87	0.87	0.6	0.67	2.2	1.14	1.21	0.87	1.44	0.79	0.63	0.91
W	0.57	0	0.59	0.95	0.33	1.35	0	0.54	0.77	0.31	0.63	0.72	0.83	0.92	0.68	1.02	0.78	1.24	2.96	0.41
Y	0.73	0.87	1.02	1.48	0.63	0.83	0.47	0.44	1.46	0.82	0.26	0.49	1.74	1.13	1.1	1.04	1.31	0.77	0.98	0.53

Amino acids in the rows precede amino acids in the columns. Propensities >1.30 are highlighted. Sample size is 14.2 ± 9.6 per residue pair.

Table VI. Amino acid pair propensities for helical linkers

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.87	0.63	0.83	1.48	0.9	0.78	1.03	0.98	1.16	1.11	1.5	1.12	0.59	1.48	1.49	0.99	0.7	1.05	1.56	1.39
C	0.98	0	0.19	0.8	0.89	0.41	1.53	0.42	0.65	0.51	2.61	0.69	0	0.64	0.31	0	0.73	0.36	0	0.34
D	0.69	0.86	0.81	1.04	0.8	0.3	1.12	0.84	1.14	1.15	0.62	1.17	1.11	1.49	1.6	0.98	0.85	0.83	1.83	0.9
E	1.63	0.33	0.76	1.38	1.32	0.83	1.49	1.08	1.21	1.52	1.52	0.96	0.62	1.71	1.75	1.04	1.51	0.99	0.92	1.05
F	0.96	0.31	0.69	1.28	1.68	0.68	0.58	1.55	0.82	0.98	1.17	1.45	0.98	1.62	0.77	1.31	0.53	1.15	1.45	1.66
G	0.88	0.56	0.61	0.55	0.88	0.42	0.66	0.59	0.62	0.9	1.23	0.74	0.65	1.21	0.84	0.65	0.64	0.6	0.39	0.9
H	0.91	0.74	0.74	1.33	0.53	0.46	1.4	0.41	0.76	1.16	1.02	0.78	1.45	0.55	1.85	0.82	1.63	1.24	0.95	1.54
I	0.73	0.92	0.91	1.36	0.98	0.57	0.75	0.87	0.79	1.12	0.59	1.08	0.58	1.21	1.17	0.72	0.74	0.58	0.61	0.63
K	0.89	0.73	0.95	1.38	0.81	0.59	1.4	1.02	0.75	1.37	1.28	1	0.94	1.42	1.59	0.94	0.77	0.89	1.06	1.4
L	1.64	1	1.24	1.64	1.75	0.86	1.47	0.84	1.2	1.26	0.88	1.26	1.22	1.23	1.44	1.03	0.73	1.08	1.42	0.96
M	1.66	0	1.26	0.81	1.3	0.95	0.39	1.1	1.2	1.5	0.7	0.54	1.64	1.08	1.79	0.63	1.12	1.08	1.62	1.52
N	1.16	0.68	1.28	1.32	0.99	0.45	1.09	1.11	1.22	1.06	0.62	0.78	0.64	1.13	0.81	0.62	0.43	1.22	1.39	0.99
P	0.65	0.25	0.86	0.95	0.9	0.72	0.72	0.43	0.56	0.93	1.31	0.72	0.36	0.38	0.83	1.16	0.81	0.82	0.83	1.13
Q	0.9	0.22	1.37	1.49	1.58	1.19	0.94	0.97	1.38	1.21	1.36	0.9	0.61	1.02	2.08	1.29	1	1.14	1.05	0.86
R	1.02	0.55	1.71	1.4	1.3	1.16	1.06	1.29	2.22	1.49	2.37	0.94	0.61	1.64	1.25	0.8	0.96	1.12	1.05	1.38
S	0.86	0.88	0.98	1.32	0.81	0.64	0.62	0.76	0.93	1.28	0.63	0.64	0.75	0.81	0.89	0.94	0.9	0.88	0.97	0.7
T	1.42	0.16	0.9	1.07	0.68	0.64	0.81	0.49	0.65	1.05	1.32	0.72	0.82	0.61	1.1	0.79	0.85	0.61	0.77	1.01
V	1.08	0.29	0.85	0.97	0.8	0.64	1.07	0.73	1.19	1.37	1.2	0.96	0.8	0.82	0.99	0.87	0.56	0.57	0.44	1.19
W	0.37	0.88	0.42	0.91	0.69	1.47	0.95	0.29	1.17	0.99	1.29	1.27	0.36	0.23	1.58	1.12	1.18	0.88	0	0.82
Y	1.14	0.54	0.88	0.96	1.26	1.23	1.4	1.56	0.86	1.53	1.37	0.7	0.83	1.81	1.36	0.75	0.94	0.87	0.96	0.65

Amino acids in the rows precede amino acids in the columns. Propensities > 1.30 are highlighted. Sample size is 14.3 ± 10.9 per residue pair.

the disruption of a pseudo-double bond. The activation energy barrier for *cis–trans* isomerization of Xaa–Pro peptide bonds has been well characterized with values ranging from 80 to 100 kJ mol^{-1} (Reimer *et al*., 1998), confirming the widely accepted notion that *cis–trans* isomerization of Xaa–Pro causes slow protein-folding rates (Brandts *et al*., 1975).

The preceding residue type, Xaa, greatly affects the Xaa–Pro *cis–trans* inter-conversion rates, e.g. when Tyr precedes Pro, it has the highest propensity to be *cis* when compared to all other amino acids (MacArthur and Thornton, 1991). Also, amino acids with aromatic side chains, Tyr, His, Phe and Trp, have been found to specifically reduce *cis–trans* isomerization rates when in the Xaa position (Reimer *et al*., 1998) and therefore slows down protein folding. Interestingly, these four residue types are the least preferred to precede Pro within both medium sized and non-helical linkers (Tables IV and V, respectively), whereas His is one of the most preferred types to precede Pro in the helical set (Table VI). The rate constants for *trans* to *cis* isomerization as measured for Ac–Ala–Xaa–Pro–Ala–Lys–NH_2 pentapeptides (Reimer *et al*., 1998) are negatively correlated with the Xaa–Pro pair linker propensities derived from our database, with a correlation coefficient of –0.54 (Figure 3). The faster the conversion from *trans* to *cis* in Xaa–Pro, the less likely the pair is to occur within a linker. Although, the reverse *cis* to *trans* conversion rate, has a smaller negative correlation coefficient of only –0.31. The pattern for a residue preceding Pro within a linker is that it will have a low propensity to be in *cis* conformation, will have slow *trans* to *cis* conversion rates and faster *cis* to *trans* rates. This again supports the notion that linkers prefer an extended conformation and act as spacers to allow domains to fold independently.

Linker database

We have constructed a database of inter-domain linkers, providing an ample source of potential linkers for novel fusion proteins. These linkers provide the conformation, flexibility and stability needed for a protein's biological function in its natural environment. The linker database is available at http://mathbio.nimr.mrc.ac.uk. The database is organized into two

Table VII. Dyad sequence codes for linkers

All linkers	Helical linkers	Non-helical linkers
Cys-Met	Cys-Met	Met-His
Cys-His	Asp-Trp	Phe-Trp
Phe-Trp	Ala-Trp	Pro-His
Pro-Trp	Trp-Gly	Tyr-Glu
His-Arg	Tyr-Gln	His-Lys
Asp-His	Tyr-Ile	Pro-Trp
Gln-Ser	His-Thr	His-Phe
Trp-Gly	Cys-His	Ile-Met
Ala-Gln	Phe-Trp	Asp-His
Met-His	Leu-Phe	Val-His
Gly-Met	His-Pro	Asp-Arg
Lys-Pro	Thr-Ala	Arg-Ile
Leu-Phe	Glu-Met	Gln-Ser
Tyr-Gln	Lys-His	Phe-Cys
Arg-Met	Met-Leu	Gly-Met
Arg-Ile	Ala-Gln	His-Arg
Phe-Tyr	Tyr-Leu	His-Asn
Phe-Asn	Phe-Ile	Glu-Cys
Thr-Ala	Lys-Tyr	Val-Thr
His-Tyr	Trp-Arg	Thr-His
Gln-Val	Leu-Ala	Cys-Arg
	Phe-Ser	Phe-Gln
	Ala-Arg	Glu-Val
	Phe-Asn	Lys-Thr
	Glu-Thr	Thr-Asp
	Leu-Trp	Ser-Phe
	Asn-Glu	Ser-Thr
	Leu-His	Gln-Val
		Tyr-Lys
		Asn-Gln
		Gln-Lys

Amino acid pairs are in order of propensity difference between the pair and their reverse complement, starting with the largest propensity. The amino acids in the columns have a linker propensity >1.3, whereas the linker propensity of the respective pair is <1.2 and the difference between the two propensities is >0.3 (values used by Crasto and Feng, 2001).

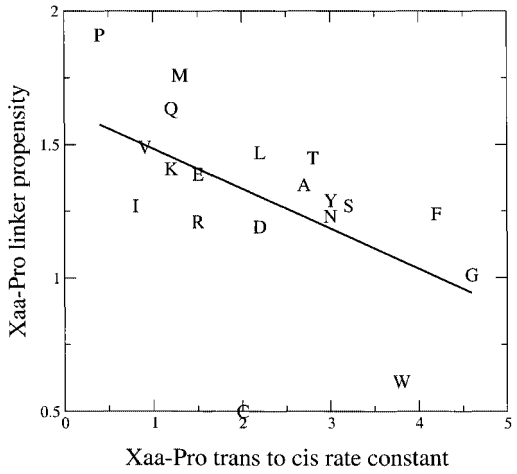

Fig. 3. Xaa–Pro linker propensities versus *trans–cis* rate constants described by Reimer *et al.* (Reimer *et al.*, 1998). Rate constants were measured at 4°C in sodium phosphate buffer (pH 5). Values for His are not included. The data shows a correlation coefficient of –0.537. Linker pair propensities are for the complete linker set.

basic classes 'non-helical' and 'helical'. The database can be searched using several query types, such as PDB code, PDB header, linker length, C_α extent or sequence. Searches using

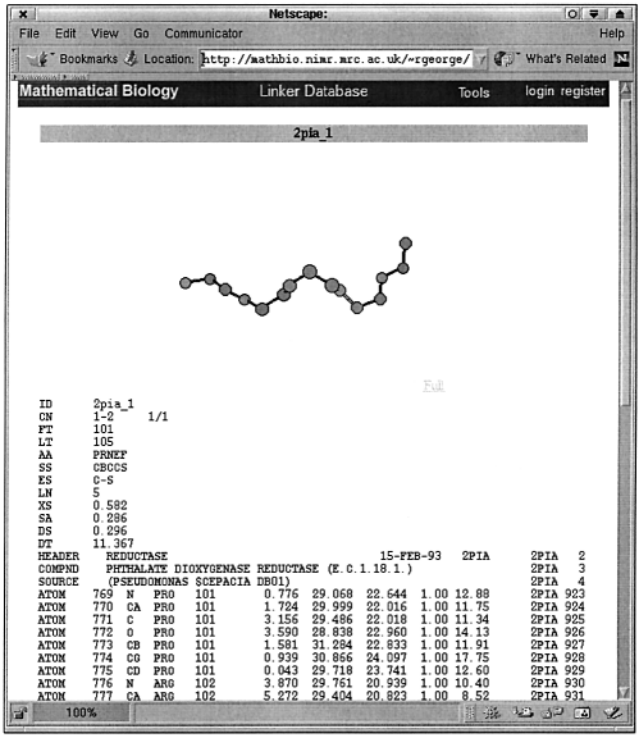

Fig. 4. Linker database output file for linker 1befA_1 (http://mathbio.nimr.mrc.ac.uk). Visualization of the linker in 3D is achieved using Jmol molecule viewer (http://jmol.sourceforge.net). The database will provide an ample source of potential linkers for novel fusion proteins.

regular expressions are possible and can be used to find particular sequence motifs. The two linker categories may be queried individually or together. The output from a search will display a list of linker identifiers along with their sequences. A hyper-link connects each linker identifier to an atomic coordinate page, which also contains an interactive rotating 3D atomic structure of the linker (Figure 4).

Conclusion

Multi-functional enzymes generally comprise a number of discrete domains connected by inter-domain linkers. The recurrent use of protein domains in the creation of novel protein and enzymic activities offers virtually unlimited possibilities (Nixon *et al.*, 1997). Inter-domain linkers are likely to facilitate the folding of multidomain proteins: α-helical linkers are thought to act as rigid spacers to prevent non-native interactions between domains that may interfere with correct domain folding. The requirement for stiffness turns out to be important also for the non-helical linkers: having a large proportion of proline residues leads to a rigidity of the polypeptide. Compared to their helical counterparts, non-helical linkers have evolved further away from intra-domain segments, to obtain the required rigidity inherent to the α-helical conformation.

However, prolines can form tight turns or *cis*-Pro isomers, which negatively affect domain independence. The chance of these conformations can be reduced by careful selection of the preceding residue, and this study confirms that this happens in nature.

The linker database introduced here, along with its query protocol, should be helpful as an initial reference for any domain fusion protocol. Linker design is an obvious necessity

in protein engineering, as they should keep domains apart while allowing them to move as part of their catalytic function. The observed natural tendency to form rigid linkers might also be related to avoiding proteolytic cleavage, as linkers are likely targets for protease degradation (Hellebust *et al*., 1989).

References

Argos,P. (1990) *J. Mol. Biol*., **211**, 943–958.

Aurora,R., Creamer,T.P., Srinivasan,R. and Rose,G.D. (1997) *J. Biol. Chem*., **272**, 1413–1416.

Barlow,D.J. and Thornton,J.M. (1988) *J. Mol. Biol*., **201**, 601–619.

Brandts,J.F., Halvorson,H.R. and Brennan,M. (1975) *Biochemistry*, **14**, 4953–4963.

Briggs,S.D. and Smithgall,T.E. (1999) *J. Biol. Chem.*, **274**, 26579–26583.

Chakrabarti,P. and Chakrabarti,S. (1998) *J. Mol. Biol*., **284**, 867–873.

Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–12.

Chou,P.Y. and Fasman,G.D. (1974) *Biochemistry*, **13**, 211–222.

Crasto,C.J. and Feng,J. (2001) *Proteins*, **42**, 399–413.

Dieckmann,R., Pavela-Vrancic,M., von Dohren,H. and Kleinkauf,H. (1999) *J. Mol. Biol*., **288**, 129–140.

Eisenberg,D. (1984) *Annu. Rev. Biochem*., **53**, 595–623.

Gokhale,R.S. and Khosla,C. (2000) *Curr. Opin. Chem. Biol*., **4**, 22–27.

Gokhale,R.S., Tsuji,S.Y., Cane,D.E. and Khosla,C. (1999) *Science*, **284**, 482–485.

Hellebust,H., Murby,M., Abrahmsen,L., Uhlen,M. and Enfors,S. (1989) *Biotechnology*, **7**, 165–168.

Ikebe,M., Kambara,T., Stafford,W.F., Sata,M., Katayama,E. and Ikebe,R. (1998) *J. Biol. Chem.*, **273**, 17702–17707.

Kabsch,W. and Sander,C. (1983) *FEBS Lett*., **155**, 179–182.

LaFevre-Bernt,M., Sicheri,F., Pico,A., Porter,M., Kuriyan,J. and Miller,W.T. (1998) *J. Biol. Chem*., **273**, 32129–32134.

MacArthur,M.W. and Thornton,J.M. (1991) *J. Mol. Biol*., **218**, 397–412.

Matsuo,Y. and Bryant,S.H. (1999) *Proteins*, **32**, 70–79.

Nixon,A.E., Warren,M.S. and Benkovic,S.J. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 1069–1073.

Ostermeier,M. and Benkovic,S.J. (2000) *Adv. Protein. Chem*., **55**, 29–77.

Packman,L.C. and Perham,R.N. (1987) *Biochem. J*., **242**, 531–538.

Pastore,A., Harvey,T.S., Dempsey,C.E. and Campbell,I.D. (1989) *Eur. Biophys. J*., **16**, 363–367.

Perham,R.N. (1991) *Biochemistry*, **30**, 8501–8512.

Radford,S.E., Laue,E.D., Perham,R.N., Miles,J.S. and Guest,J.R. (1987) *Biochem. J.*, **247**, 641–649.

Radford,S.E., Laue,E.D., Perham,R.N., Martin,S.R. and Appella,E. (1989) *J. Biol. Chem*., **264**, 767–775.

Reimer,U., Scherer,G., Drewello,M., Kruber,S., Schutkowski,M. and Fischer,G. (1998) *J. Mol. Biol.*, **279**, 449–460.

Robinson,C.R. and Sauer,R.T. (1998) *Proc. Natl Acad. Sci.* USA., **95**, 5929–5934.

Russell,G.C. and Guest,J.R. (1991) *Biochim. Biophys. Acta*, **1076**, 225–232.

Taylor,W.R. (1999) *Protein Eng*., **12**,203–216.

Tieleman,D.P., Shrivastava,I.H., Ulmschneider,M.R. and Sansom,M.S. (2001) *Proteins*, **44**, 63–72.

van Leeuwen,H.C., Strating,M.J., Rensen,M., deLaat,W. and van der Vliet,P.C. (1997) *EMBO J.*, **16**, 2043–2053.

Wootton,J.C. and Drummond,M.H. (1989) *Protein Eng.*, **2**, 535–543.

Received June 25, 2002; accepted August 26, 2002