Protein Engineering vol.14 no.6 pp.397–402, 2001

A critical view on conservative mutations

Per Harald Jonson and Steffen B.Petersen[1]

Biostructure and Protein Engineering Group, Department of Life Sciences, Aalborg University, Sohngaardsholmsvej 49, DK-9000 Aalborg, Denmark

[1]To whom correspondence should be addressed. E-mail: sp@bio.auc.dk

By analysing the surface composition of a set of protein 3D structures, complemented with predicted surface compositional information for homologous proteins, we have found significant evidence for a layer composition of protein structures. In the innermost and outermost parts of proteins there is a net negative charge, while the middle has a net positive charge. In addition, our findings indicate that the concept of conservative mutation needs substantial revision, e.g. very different spatial preferences were found for glutamic acid and aspartic acid. The alanine screening often used in protein engineering projects involves the substitution of residues to alanine, based on the assumption that alanine is a 'neutral' residue. However, alanine has a high negative correlation with all but the non-polar residues. We therefore propose the use of, for example, serine as a substitute for the residues that are negatively correlated with alanine.

Keywords: amino acid properties/protein engineering/solvent accessibility/spatial contacts/structural preference

Introduction

Upon folding of a peptide chain into a 3D protein structure, some residues are transferred from a polar environment to a more non-polar environment in the interior of the folded protein. This transfer is driven by the thermodynamic properties of the amino acids and the solvent. Throughout molecular evolution nature has selected for suitable function and stability of the resulting protein. For small to medium sized proteins—in the folded structure—only a few residues are totally buried (Chothia, 1976; Miller *et al*., 1987; Petersen *et al*., 1998), whereas most residues are only partially buried. The variation in solvent accessibility is dependent on the properties of the residue in question and is reflected in the amino acid composition throughout the protein structure. These differences in the solvent accessibility profile have found wide applications in various structure prediction methods (Holbrook *et al*., 1990; Rost and Sander, 1994; Thompson and Goldstein, 1996). Also, the use of environment specific substitution matrices (Donnelly *et al*., 1994; Wako and Blundell, 1994) have proven valuable. The sequential neighbourhood of amino acids has been investigated previously (Vonderviszt *et al*., 1986) and its use has been found in, for example, loop prediction (Wojcik *et al*., 1999) and secondary structure prediction (Chou and Fasman, 1978; Chandonia and Karplus, 1999; Jones, 1999). No significant correlation between residues sequential neighbour preference was discovered.

The spatial neighbourhood around individual residues has also been previously investigated (Burley and Petsko, 1985; Bryant and Amzel, 1987; Miyazawa and Jernigan, 1993; Petersen *et al*., 1999). Further, spatial contacts have been studied to derive contact potentials for the different amino acid interactions (Brocchieri and Karlin, 1995; Miyazawa and Jernigan, 1996, 1999). The common strategy is to study the number of contacts within a given distance cut-off. However, the literature seems devoid of investigations of distance-dependent contacts and also of reports utilizing the embedded information of the solvent accessibility of the residues involved.

A two-state prediction of solvent accessibility correlation between hydrophobicity, buried contact propensity and the location in the prediction window has been reported (Mucchielli-Giorgi *et al*., 1999). However, it does not describe any correlation between individual residue distributions.

It is important to be able to discriminate between correctly folded and misfolded model structures. It has been pointed out that potential energy-based methods do not discriminate well between folded and misfolded structures. However, structural features such as buried polar surface (Overington *et al*., 1992) and number of polar contacts (Bryant and Amzel, 1987; Golovanov *et al*., 1999) have proven valuable.

In protein engineering the concept of conservative mutations is frequently used. The general idea is that a substitution of an amino acid with another amino acid with similar physicochemical properties will not influence the stability and function of the protein. The present paper shows that the spatial preferences for similar residues can be dramatically different in protein structures under similar circumstances (in this context solvent accessibility).

The results of the neighbour analysis will be valuable in model validation, as a tool for structure prediction and especially as a guide in the search for stability enhancing mutations.

Methods

The sequences used are a subset of the 25% sequence identity set of non-homologous structures (Hobohm *et al*., 1992; Hobohm and Sander, 1994) derived from the protein structure databank PDB (Bernstein *et al*., 1977). Only single-chain protein sequences were used. The resulting dataset consisted of 336 single-chain sequences with a maximum pairwise sequence identity of 25%. The subset was expanded through the use of the corresponding HSSP-files (Dodge *et al*., 1998). The total data set contained 8379 aligned sequences and 1 415 986 residues. This corresponds to 6.7% of all residues in version 34 of SWISS-PROT (Bairoch and Apweiler, 1997). The length of the sequences was between 64 and 1017 residues. The resolution of the X-ray structures used varied between 1.0 and 3.0 Å, with an average of 2.0 Å. Further, the subset contained 31 structures solved by NMR. However, all hydrogen-atom co-ordinates were discarded. To check for a possible bias introduced by the use of the homologous sequences the complete analysis was done with and without the aligned sequences. No significant differences were observed, although the reduced size of the smaller of the two datasets, as expected, gave rise to more noise.

The spatial neighbours of each residue were determined based on solvent accessibility and spatial distance. The solvent accessibility was taken from the respective HSSP-files (Dodge *et al.*, 1998). For each surface residue the neighbouring surface residues were grouped according to their distance to the residue in question. The distance between two residues was computed as the shortest distance among the set of all possible pairs of atoms in the two residues. We assume that the alignment in the HSSP-file implies that neighbours in the main sequence are also neighbours in the aligned sequences and that the solvent accessibility is conserved (Andrade *et al.*, 1998; Goldman *et al.*, 1998). The expected number of neighbour interactions between residues of type i and j are calculated by

$$N^{\text{expected}}_{ij|d,ACC} = x_{j|d,ACC} \cdot x_{j|d,ACC} \cdot N_{0|d,ACC} \qquad (1)$$

where x_i and x_j are the fraction of amino acid i and j in the dataset for the distance range d and at a solvent accessibility larger than the cut-off ACC and N_0 is the total number of observed neighbour contacts. The score, $S_{ij|d,ACC}$, is calculated by

$$S_{ij|d,ACC} = \ln\left(N^{\text{observed}}_{ij|d,ACC} \,/\, {}^{\text{expected}}_{ij|d,ACC}\right) \qquad (2)$$

This gives a negative score for disfavoured neighbour-pairs and a positive score for favoured interactions. The score value $S_{ij|d,ACC}$ can be transformed into an apparent thermodynamic parameter by multiplication with RT.

The net charge in each layer of the protein was calculated. Aspartic acid and glutamic acid are considered negatively charged and arginine and lysine are considered positively charged. Histidine is either considered as uncharged or positively charged. The relative net charge, Δq_{rel}, we define as

$$\Delta q_{\text{rel}} = (N_{\text{Positive}} - N_{\text{Negative}}) \,/\, N_{\text{Total}} \qquad (3)$$

where N_{Positive} is the number of positive residues, N_{Negative} the number of negative residues and N_{Total} the total number of residues in that particular layer.

The PDB identification codes for the structures used are 1ptx, 2bbi, 1hcp, 1iml, 1cdq, 1vcc, 1nkl, 1tiv, 2abd, 2hts, 1tpg, 1fbr, 1pco, 1who, 1beo, 2ncm, 1fim, 1tlk, 1xer, 1onc, 1rga, 1erw, 1fd2, 1put, 1fkj, 1jpc, 1thx, 1jer, 1ccr, 1wad, 2tgi, 1pls, 1neu, 4rhn, 1rmd, 1hce, 1hfh, 1tam, 2pf1, 1bip, 1whi, 1yua, 1bp2, 1zia, 4fgf, 7rsa, 1bw4, 2vil, 1eal, 1rie, 1doi, 3chy, 1cpq, 1msc, 1mut, 1rcb, 1lzr, 1htp, 1lid, 1lis, 1lit, 1kuh, 1nfn, 1irl, 1poc, 2tbd, 1cof, 1pms, 1rsy, 1snc, 1eca, 1jvr, 2end, 1anu, 5nul, 1fil, 1jon, 1lcl, 1itg, 1tfe, 1maz, 1pkp, 1lba, 1vsd, 2fal, 1ash, 1def, 2hbg, 1div, 1gds, 1grj, 1i1b, 1ilk, 1rcy, 1sra, 1ulp, 1mbd, 1aep, 1jcv, 2gdm, 1phr, 1rbu, 1esl, 1hlb, 1mup, 1vhh, 1gpr, 1btv, 1cyw, 1klo, 1l68, 3dfr, 2cpl, 1sfe, 1huw, 5p21, 1ha1, 1wba, 1lki, 2fha, 1prr, 2fcr, 1amm, 1cid, 1hbq, 1cdy, 2stv, 153l, 1rec, 1xnb, 2sas, 1gky, 1knb, 1ryt, 1zxq, 1har, 1cex, 1chd, 2tct, 2ull, 1gen, 1iae, 1nox, 1rnl, 2gsq, 1cfb, 1dyr, 1nsj, 2hft, 1fua, 2eng, 1thv, 1hxn, 2abk, 9pap, 1lbu, 3cla, 1vid, 2ayh, 2dtr, 1gpc, 1dts, 1jud, 1emk, 1ois, 1akz, 1sgt, 1ad2, 1nfp, 1din, 1lrv, 1dhr, 1bec, 1lbd, 1dpb, 1jul, 1mrj, 1fib, 1hcz, 1mml, 1vin, 1dja, 2cba, 3dni, 1lxa, 1arb, 1rgs, 1tys, 3tgl, 1ako, 1eny, 1ndh, 2dri, 1xjo, 1drw, 1kxu, 2prk, 1cnv, 1tfr, 1ytw, 1iol, 2ebn, 1tml, 1han, 1xsm, 1pbn, 1amp, 1ryc, 1bia, 1vpt, 1csn, 2ora, 1ctt, 1bco, 1fnc, 1gym, 1pda, 1cpo, 1esc, 2reb, 1mla, 1sig, 8abp, 1ghr, 1iow, 2ctc, 1gca, 1sbp, 1ede, 1pgs, 2cmd, 1anv, 1gsa, 1tag, 1dsn, 2acq, 1cvl, 1tca, 2abh, 2pia, 1pot, 1vdc, 1axn, 1msk, 1hmy, 2bgu, 1ldm, 1dxy, 1ceo, 1nif, 1arv, 1xel, 1uxy, 1rpa, 2lbp, 3pte, 1uby, 1fkx, 1pax, 3bcl, 1air, 1mpp, 2mnr, 1eur, 1cem, 1fnf, 1pea, 1omp, 2chr, 1pud, 1kaz, 1mxa, 1edg, 2sil, 1ivd, 1pbe, 1svb, 1ars, 1oyc, 1inp, 1oxa, 1eft, 1phg, 1cpt, 1iso, 1qpg, 2amg, 1uae, 1gnd, 2dkb, 1gpl, 1csh, 4enl, 1pmi, 1lgr, 1nhp, 1gcb, 1bp1, 1geo, 2bnh, 3grs, 1gln, 1gai, 2pgd, 2cae, 2aaa, 1byb, 1smd, 2myr, 3cox, 1dpe, 1pkm, 1ayl, 1crl, 1ctn, 1clc, 1tyv, 2cas, 1ecl, 1oxy, 1vnc, 1gal, 1dlc, 1sly, 1dar, 1gof, 1bgw, 1aa6, 1vom, 8acn, 1kit, 1taq, 1gpb, 1qba, 1alo and 1kcw.

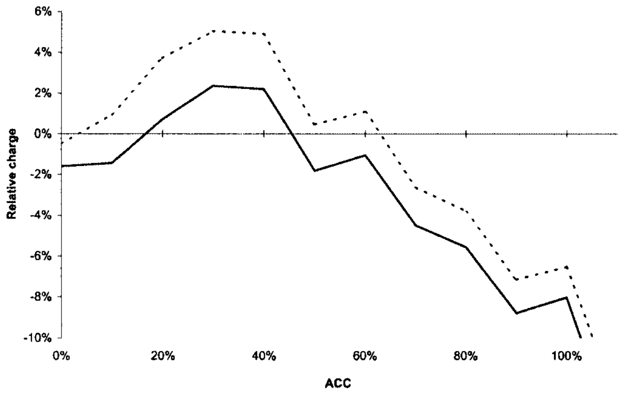

Fig. 1. The net relative charge in layers of protein structures with different solvent accessibility (ACC). The net relative charge is defined as the net charge per residue found in a particular layer (number of positive charges – number of negative charges/number of residues). Aspartic acid and glutamic acid are considered to be negative, arginine, lysine and protonated histidine positive (dotted line). The solid line include all of the above-mentioned residues except histidine.

Results and discussion

The distribution of charged residues in different layers of the protein 3D structure and the total net charge are shown in Figure 1. In the innermost and outermost parts of proteins there is a net negative charge, while the middle has a net positive charge. This apparent three layer structure with alternating charge exposing the negatively charged outermost layer to the solvent is interesting. Such organisation will secure some level of radial charge neutralisation, and may possibly contribute to tight packing of the protein. Likewise this charge organisation of the surface layer could provide important electrostatic guidance during the folding event. Conversely, changing pH to acidic or alkaline conditions at which subsets of the titratable residues becomes uncharged will destabilise the packing of residues at the surface of the protein. Buried, acidic amino acids can be found in several different protein structures and these residues play important functional roles in, for example, trypsin (McGrath *et al.*, 1992), ribonuclease T1 (Giletto and Pace, 1999) and thioredoxin (Dyson *et al.*, 1997; Bhavnani *et al.*, 2000). The reported three layer structure is observed both with and without the aligned sequence and is therefore not caused by a bias introduced by the conservation of the buried, charged groups within a protein family.

The spatial neighbours around each type of residue were calculated without any discrimination for solvent accessibility. With the notable exceptions of tryptophan and cysteine, amino acids were not frequently observed as spatial neighbours to identical residue types. This trend was not dependent upon the choice of distance cut-off (results not shown). The differences in distribution were remarkably small between the different

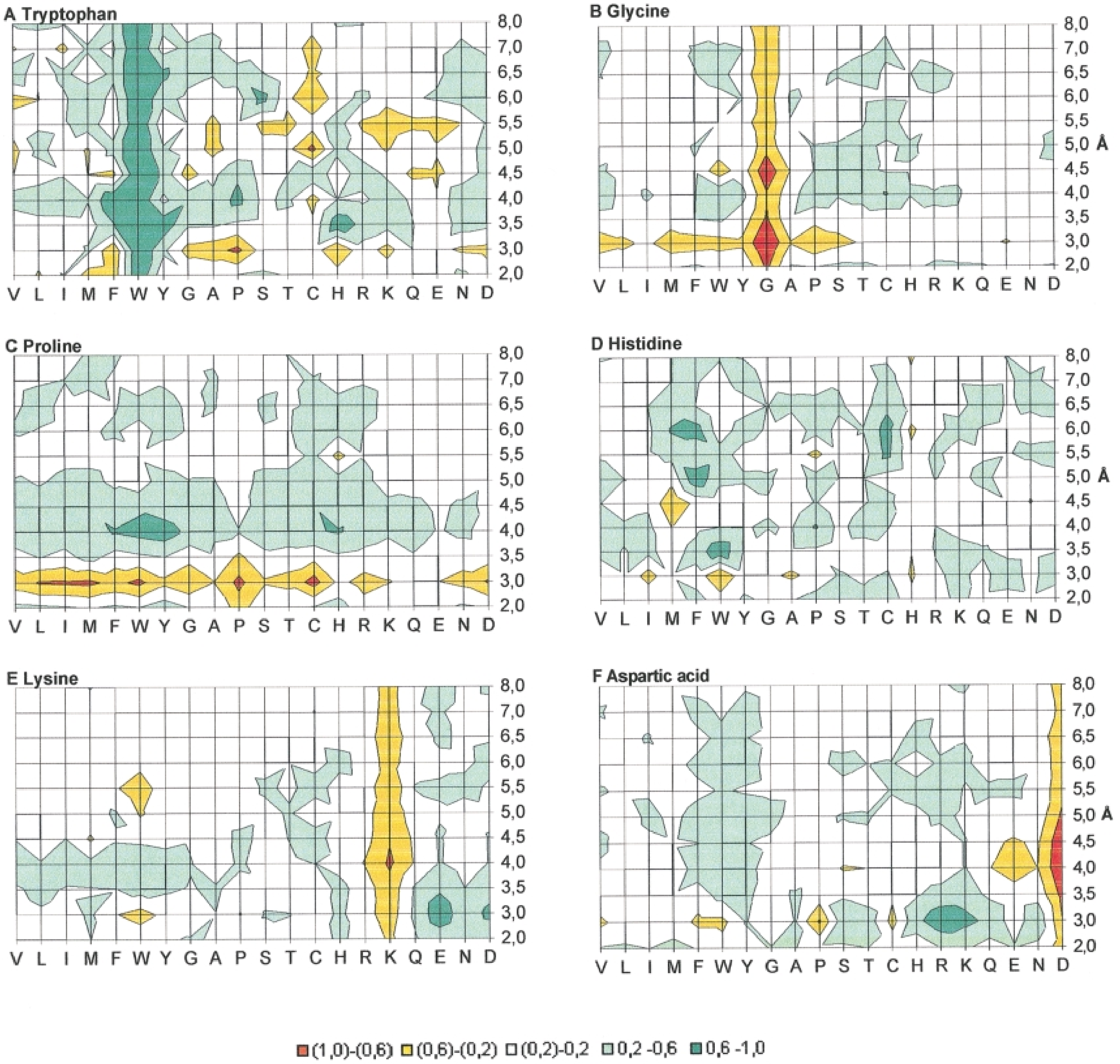

Fig. 2. The significantly over- or under-represented neighbour pairs. All residues have a solvent accessibility higher than 20%. The distance in Å between the residues is given along the vertical axis. Red and green represent areas where the number of pairs are less and higher than expected, respectively. (A) Tryptophan; (B) glycine; (C) proline; (D) histidine; (E) lysine; (F) aspartic acid.

amino acids for an 8 Å distance cut-off suggesting that 8 Å is a large enough distance for the distribution to become independent of the nature of the central residue. This observation led to the use of 8 Å as the largest distance between neighbours investigated in detail.

Figure 2 shows the score values for all amino acid neighbour pairs involving tryptophan, glycine, alanine, proline, serine, histidine, lysine and aspartic acid for neighbour pairs with at least 20% solvent accessibility. The results for the other amino acids are available on our homepage (http://www.bio.auc.dk/). Score values have been calculated similarly for other solvent accessibility cut-offs. The aromatic residue tryptophan is one of only two residues showing a clear preference for contacts with the same residue type (the other is cysteine). Also interactions with the other aromatic residues are preferred. Interestingly the interactions between tryptophan and the two acidic residues (aspartic acid and glutamic acid) seem different. While tryptophan and glutamic acid are observed less frequently than expected, the opposite is observed for tryptophan and aspartic acid. Glycine shows the typical negative score for interactions with the same residue type. Also, glycine does not seem to have neighbours in the close spatial neighbourhood (≤3.5 Å). This under-representation of neighbours close by is even clearer for proline. We interpret this under-representation as a sign of the preference for loop that proline residues have. The lack of interactions with all other amino acids in its

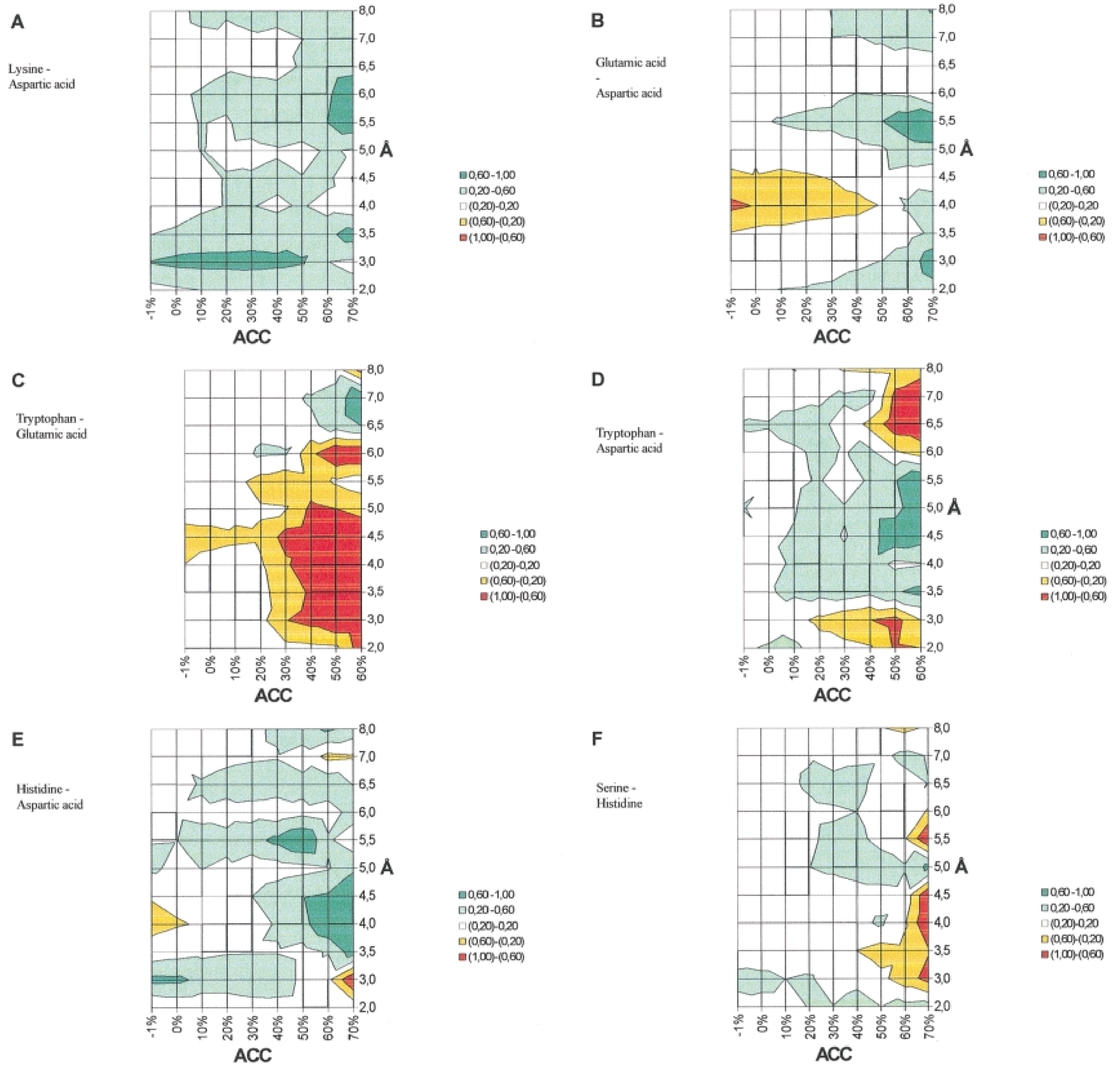

Fig. 3. Over- and under-represented neighbour pairs as a function of solvent accessibility (ACC) and distance (Å). The distance between the residues is given along the vertical axis and the solvent accessibility along the horizontal axis. (A) Lysine–aspartic acid; (B) glutamic acid–aspartic acid; (C) tryptophan–glutamic acid; (D) tryptophan–aspartic acid; (E) histidine–aspartic acid; (F) serine–histidine.

vicinity point to most contacts being with solvent molecules. However, proline has an abundance of contacts at a larger distance (4–5 Å). Histidine is interesting in that it shows signs of its aromatic properties, through preference for contacts with aromatic residues (~3.5 Å), and its polarisable nature, through preferred contacts with the negatively charged residues (~3 Å). The basic amino acid lysine has as expected a clear negative score for contacts with other lysines. The favourable electrostatic interactions with the acidic amino acids is evident.

Some of the most interesting pair interactions are shown in Figure 3. Figure 3A depicts the salt bridge pair lysine–aspartic acid. The strong over-representation seen at 3 Å separation is consistent with the classical salt bridge concept. The over-representation of lysine–aspartic acid pairs in the most solvent exposed layers observed at 5.5 to 6 Å is unexpected. We propose that charge networks on the protein surface could cause this observation. In Figure 3B the result for the glutamic acid–aspartic acid pair is shown. The most obvious feature is the expected under-representation of this pair. However, close to the protein surface the same restriction does not appear to be present. Again we propose that surface located charge networks are contributing to this observation. In Figures 3C and D the amino acid pairs tryptophan–glutamic acid and tryptophan–aspartic acid are shown. The common belief that a glutamic acid to aspartic acid mutation is conservative is contrary to the observations shown. The tryptophan–glutamic acid pair is highly under-represented in the highly solvent exposed layers of the proteins. Surprisingly, the same cannot be said for the tryptophan–aspartic acid pair, where an over-representation is observed for the 3.5 to 6 Å distance interval. Similar, but less pronounced, observation was made for the tyrosine–glutamic acid and tyrosine–aspartic acid pairs. No significant differences were observed between the phenylalanine–glutamic acid and phenylalanine–aspartic acid pairs. The only difference between the glutamic acid and the aspartic acid is the length of the side chain. Common to both tryptophan and tyrosine is their polarisability, in contrast to phenylalanine. We believe that surface located tryptophans involved in defining protein functionality are polarised by their local electrostatic environment. Although we cannot provide a quantitative explanation, it is plausible that the differences between the different chain length of glutamic acid and aspartic acid may put preference on the proximity to tryptophan. It has been shown that aspartic acid has a tendency to have favourable interactions between the side chain carbonyl group and the backbone carbonyl group (Deane *et al*., 1999), resulting in a ring-like structure. Similar conformations have not been observed for glutamic acid. In Figures 3E and F the histidine–aspartic acid and serine–histidine pairs are shown. Since these three residues constitute the active site residues of a wide range of hydrolases they have particular interest. There is an over-representation of histidine–aspartic acid pairs in the highly solvent accessible areas. The distance is larger than the typical distance observed in active site crevasses. However, the small, but significant, over-representation in the 3 Å range conforms with the classical histidine–aspartic acid distances in hydrolases. Figure 3E shows the clear preference for contacts between buried histidines and aspartic acids. We believe that this feature is an important part of the molecular evolution of *de novo* catalytic sites. Storing possible catalytic 'triads' in non-functional environments makes the number of amino acid substitutions necessary to activate the site smaller.

The most distinct feature in Figure 3F is the clear under-representation of serine–histidine pairs in highly solvent exposed environments. A weak over-representation of the serine–histidine pair is seen at 3 Å in the less solvent accessible areas. Thus the presence of the catalytic triad apparently is determined mostly by the preference of the histidine–aspartic acid pair although the serine–histidine pair reveal similar, but much weaker, trends.

The amino acid composition of each solvent accessibility layer was determined. As expected the buried parts of the proteins are composed of a higher amount of non-polar residues than the more solvent exposed layers. The correlation between the amino acid composition was calculated from the data of the composition of the individual structural layers. Amino acids that have similar preferences for solvent contact and local environment are expected to show a high positive correlation because of similar trends in their distribution. Hence, amino acids showing negative correlation will have different preferences for local environment and are therefore not believed to be compatible, i.e. a single site mutation of this type at this location is not recommended. As the non-polar residues are abundant in the core and show a gradual decrease as the solvent accessibility increases in general the correlation between the non-polar residues is positive (Figure 4). In contrast, the polar residues are more abundant in the highly exposed parts and hence are negatively correlated with the non-polar residues. Histidine and threonine behave markedly differently. They show positive correlation to each other, but little correlation with any of the other columns, with the exception of arginine and glycine. This is caused by the low occurrence of histidine and threonine in both the buried and highly exposed areas and their relatively high occurrence in the medium exposed layers. Histidine has positive correlation with two aromatic residues, tryptophan and tyrosine, and with the weakly polar threonine and the polar arginine. Again we interpret this as a sign of both the aromatic properties and the charge properties of histidine. The weakly polar residues do not have the same clear similarity in distribution as the polar and non-polar residues. Proline and serine seem to be more closely related to the polar residues. The weakly polar residue alanine has positive correlation only with the non-polar residues. We propose that mutations between residues with high positive correlation have a high chance of maintaining the thermodynamic stability of the 3D structure. This is particularly so for charged residues. In contrast, the residues with a high degree of negative correlation are typically residues with different physical-chemical properties, which cannot be interchanged without changing the physical chemistry of the protein. The non-correlated residues involve residues with a special role in the structure, e.g. some residues often involved in catalysis. We believe that the observation that proline in our study behaves similarly to polar residues is related with the structural role of proline residues and its preference for loops and turns. The alanine screening often used in protein engineering projects involves the substitution of residues to alanine, based on the assumption that alanine is a 'neutral' residue. However, our data shows that alanine has a high negative correlation with all but the non-polar residues. We therefore propose the use of, for example, serine as a substitute for the residues that are negatively correlated with alanine.

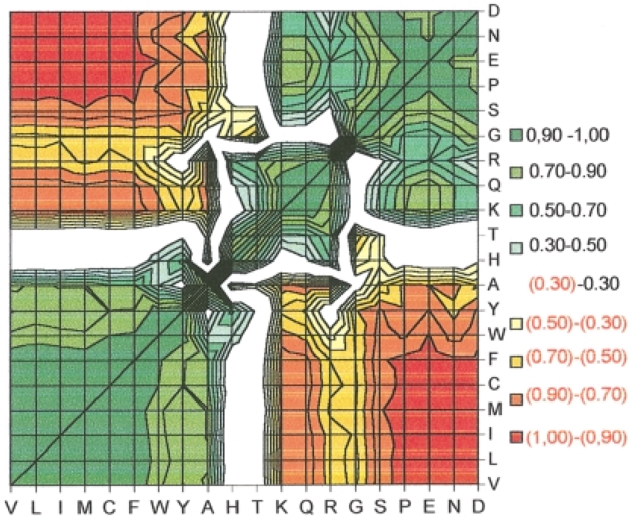

Fig. 4. Correlation between distribution of amino acids in proteins. The correlation is calculated based on the amino acid composition of the different layers of solvent accessibility layer of the protein structure. The green areas represent positive correlation, whereas the red areas represent negative correlation. Areas with low degree of correlation are in white.

In the authors' opinion the present paper provides important new information about protein structural organisation. The protein surface should be viewed as a multi-layered structural feature of the protein, where each layer has its specific composition and resulting characteristics. This simple key

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.