JMB

Adaptation of Protein Surfaces to Subcellular Location

Miguel A. Andrade[1], Seán I. O'Donoghue[2] and Burkhard Rost[2]

[1]*European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK*

[2]*European Molecular Biology Laboratory, D-69012 Heidelberg, Germany*

In vivo, proteins occur in widely different physio-chemical environments, and, from *in vitro* studies, we know that protein structure can be very sensitive to environment. However, theoretical studies of protein structure have tended to ignore this complexity. In this paper, we have approached this problem by grouping proteins by their subcellular location and looking at structural properties that are characteristic to each location. We hypothesize that, throughout evolution, each subcellular location has maintained a characteristic physio-chemical environment, and that proteins in each location have adapted to these environments. If so, we would expect that protein structures from different locations will show characteristic differences, particularly at the surface, which is directly exposed to the environment. To test this hypothesis, we have examined all eukaryotic proteins with known three-dimensional structure and for which the subcellular location is known to be either nuclear, cytoplasmic, or extracellular. In agreement with previous studies, we find that the total amino acid composition carries a signal that identifies the subcellular location. This signal was due almost entirely to the surface residues. The surface residue signal was often strong enough to accurately predict subcellular location, given only a knowledge of which residues are at the protein surface. The results suggest how the accuracy of prediction of location from sequence can be improved. We concluded that protein surfaces show adaptation to their subcellular location. The nature of these adaptations suggests several principles that proteins may have used in adapting to particular physio-chemical environments; these principles may be useful for protein design.

© 1998 Academic Press Limited

Keywords: protein evolution; protein three-dimensional-structure; protein surface; subcellular location; bioinformatics

Introduction

For over 30 years, researchers have sought to discover the principles that determine the fold of globular proteins in aqueous environments. The goals of this research effort are to predict tertiary structure from sequence, and to facilitate the design of proteins with novel structures and functions. Several general principles are now well understood: most residues with charged or polar side-chains occur at the proteins surface; most residues with apolar side-chains are buried; many of the polar chemical groups that occur in the protein interior are hydrogen bonded, effectively damping their polarity; this balance of hydrophilic and hydrophobic interactions stabilises the structure. However, these general principles are insufficient for predicting protein fold (e.g. see Rost & O'Donoghue, 1997).

One complexity in studying general properties of protein structure is that different proteins experience different physio-chemical environments, and that the exact environment influences the structure. An extreme example is globular versus transmembrane proteins. There are distinct differences in composition between these two groups of proteins that are well understood; from analysis of the sequence, it is possible to predict to which group a protein belongs at more than 97% accuracy (Rost *et al.*, 1996). Theoretical studies have tended to ignore the effect of different aqueous environments on globular proteins, although we know from *in vitro* experiments that proteins can be exquisitely sensitive to variations in pH or in the concentration of various ions. In this paper, we have approached this problem by grouping globular proteins by their subcellular location; our rationale is that all proteins in the same subcellular compartment experience a

similar physio-chemical environment, and each protein usually occurs in only one compartment. Hence, looking for protein structure properties that are characteristic to each compartment, we may uncover principles of the influence of specific environments on protein structures.

The physio-chemical environment of subcellular compartments varies with different cell types; the extracellular environment varies especially widely. However, certain features are common to almost all cell types. Compared to the extracellular environment, the cell interior has a higher overall concentration of ions, small (usually charged) metabolites, and proteins. The intracellular pH is regulated, typically to be slightly alkaline. All cells actively export specific anions (usually Na^+ in animal cells and H^+ in plant cells) to counteract osmotic expansion and to facilitate active import of specific molecules. This transport maintains a voltage difference across the plasma membrane of about -100 mV (inside more negative) and means that certain ions (H^+, Na^+) are more highly concentrated in the extracellular environment.

The physio-chemical environments of the cytoplasm and the nucleus (in eukaryotic cells) are similar, since the nuclear pore complexes are permeable to small (<5 kDa) neutral molecules (Dingwall & Laskey, 1986); however, there are differences in ionic strength due to selective ionic permeability of the nuclear envelope (Dingwall, 1991), and the large negative charge arising from the DNA phosphate backbone, which leads to accumulation of anions.

The subcellular location of eukaryotic proteins is determined by a trafficking system, which is reasonably well understood (Pfeffer & Rotteman, 1987). The system has two main branches that divide at the first stage of protein synthesis on the ribosomes: on one branch, proteins are synthesised in the cytoplasm, and from there can go to the nucleus, mitochondria, or peroxisomes; the second branch leads to the endoplasmic reticulum, then to the Golgi apparatus, and from there to lysosomes, secretory vesicles, or the cell surface (Figure 1). At each branch point in the trafficking system, a "decision" must be made for each protein; either retain the protein in the current compartment or transport it to the next compartment. These "decisions" are made by membrane transport complexes, which respond to signals on the proteins themselves (Verner & Schatz, 1988; Briggs & Gierasch, 1986; Sjöström *et al.*, 1987; von Heijne, 1985; Nielsen *et al.*, 1997). The best understood branch point is the initial division between the two main branches; proteins destined for the endoplasmic reticulum/extracellular branch have an N-terminal signal peptide that causes them to be transferred into the endoplasmic reticulum as they are being synthesised; proteins lacking this signal are synthesised in the cytoplasm. The protein signals used at the other branch points are not always so clear for two reasons: firstly, the signals are presented by folded proteins, and hence are not
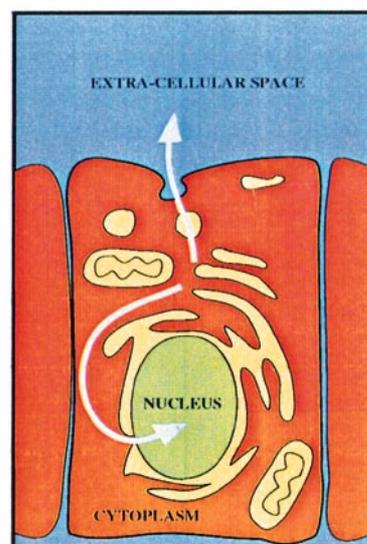

Figure 1. Subcellular locations in eukaryotic cells. In this study, we considered only proteins occurring in the three major subcellular locations: the nucleus (green), cytoplasm (red), and extracellular space (blue). Proteins from other subcellular compartments (such as the mitochondria, ribosomes, lysosomes, and the endoplasmic reticulum, all shown in yellow) were not considered. The grey arrows indicate the two major branches of the protein trafficking system that determines subcellular location: following protein synthesis in the ribosomes, one branch leads into the cytoplasm, and from there into the nucleus; the second branch leads into secretory vesicles and then to the extracellular space.

always contiguous in amino acids sequence; secondly, even where the signals are contiguous in sequence, not all of these signal peptides have been documented.

Knowing where a protein occurs in the cell is an important step towards understanding the function of the protein. Hence, a method for accurately predicting subcellular location from the amino acid sequence alone would be valuable in interpreting the wealth of data being provided by large scale sequencing projects. Indeed, we believe that prediction of physically meaningful and clearly defined quantities, such as location or secondary structure, may be more fruitful than trying to predict function directly, as "function" is necessarily very difficult to define.

To predict location, the first method would be to search for signal peptides in the sequence; unfortunately, in many cases no signal peptide can be found. Another approach would be to infer location by sequence homology to a protein with known location; however, such inference can be unreliable (see Results). A third approach was suggested by the results of Nishikawa *et al.* (1983a,b) and Nakashima & Nishikawa (1994); they found that the total amino acid composition of proteins is correlated to the subcellular location. This has recently led to a location prediction method based only on composition (Cedano *et al.*,

1997). Probably the most comprehensive location prediction method to date is the expert system developed by Nakai *et al*. (1988) and Nakai & Kanehisa (1991, 1992); this system uses a small number of rules based on composition; it is based mostly on lists of known signal peptides. No method so far has combined all three approaches; hence there is much scope for developing more accurate and general solutions to the protein location problem.

Within each subcellular compartment of a given cell type, proteins have co-evolved with the physio-chemical environment so that they are stable and functional in that environment. However, the general features of the nuclear, cytoplasmic, and extracellular environments discussed above have been constant factors throughout eukaryotic evolution. We hypothesize that these constant factors imply a set of ''environmental'' constraints on the evolution of protein structure, and that proteins will have adapted to these different environmental constraints. These environmental constraints would be distinct from the more familiar evolutionary constraints that conserve residues in the active site, residues involved in binding other macromolecules, or residues that anchor the protein within a given structural family (Rost, 1997). Rather than acting on specific residues, the effect of environmental constraints would be more global. If the hypothesis is true, we would expect distinct and measurable differences in structural properties of proteins from different compartments. The surface residues should be most affected, as they are in direct contact with the environment, whereas buried residues are largely shielded from the environment.

To test this hypothesis, we have examined all three-dimensional structures of eukaryotic proteins for which the location is annotated (as either nuclear, cytoplasmic, or extracellular). In agreement with the hypothesis, we found evidence that protein surfaces have adapted to the particular environment in each compartment.

Results

In compiling the data sets, it rapidly became clear that subcellular location is not annotated for most of the proteins in PDB, and hence the *Single* and *Glycosylated* data sets were relatively small, while the *Non-located* class was large (Table 1). In compiling the *Homology* data set, we found ten sequence families in which proteins occurred in two different subcellular locations (two families had members in both the nucleus and cytoplasm; eight had members in both the cytoplasm and extracellular space). Thus even at 40% sequence identity, it is not safe to infer that two proteins have the same subcellular location.

Figure 2(a) shows the total amino acid composition vectors of the *Single* data set projected onto the plane defined by the first two principal components. Proteins from the three location classes fall into three clusters; although there is some overlap between the clusters, the centres are distinct. As discussed in Methods, the occurrence of clusters in the principal component projection indicates that the total composition vectors are correlated with location, in agreement with the observation of Nishikawa *et al*. (1983a,b) In Figure 2(b), the plane and projections have been calculated from the surface composition vectors of the *Single* data set. Again, we observe a clustering by location class; the clusters are better defined with a larger separation of the cluster centres, and a slight decreased in overlap. In contrast, the interior composition vectors of the same data set do not form distinct clusters (Figure 2(c)). Hence we concluded that the signal observed for the total amino acid composition was due almost entirely to the surface residues.

Figure 3 shows the surface composition vectors for the *Homology* data set projected onto the plane defined by the average vectors of each of the three location classes. If we accept the hypothesis that the surface composition vectors are correlated with subcellular location, then this projection gives the optimal view to examine how uniquely defined each cluster is. In addition, since the *Homology* data set contains many more protein sequences than the *Single* data set, it provides a much more telling test of cluster overlap. In fact, proteins from the three location classes are grouped into very clearly defined clusters with little overlap between the classes (Figure 3).

Table 1. Breakdown of the data sets used in this study

Data set	Nuclear	Cytoplasmic	Extracellular	Total
Single	44	66	11	121
Glycosylated	–	–	16	16
Homology	446	885	30	1361
Non-located	?	?	?	262

For each of the data sets used in this study, the Table lists the number of proteins in the three major locations (nuclear, cytoplasm, extracellular) and the total in each data set. The *Single* data set is defined as all sequence-distinct, eukaryotic, non-glycosylated proteins in the PDB, and known to occur in one of the three locations; *Glycosylated*, as for the *Single* data set except glycosylated protein; *Homology*, all protein sequences with known location and with high homology to a structure in the *Single* data set; *Non-located*, all sequence-distinct eukaryotic structures in the PDB for which the subcellular location has not been annotated.
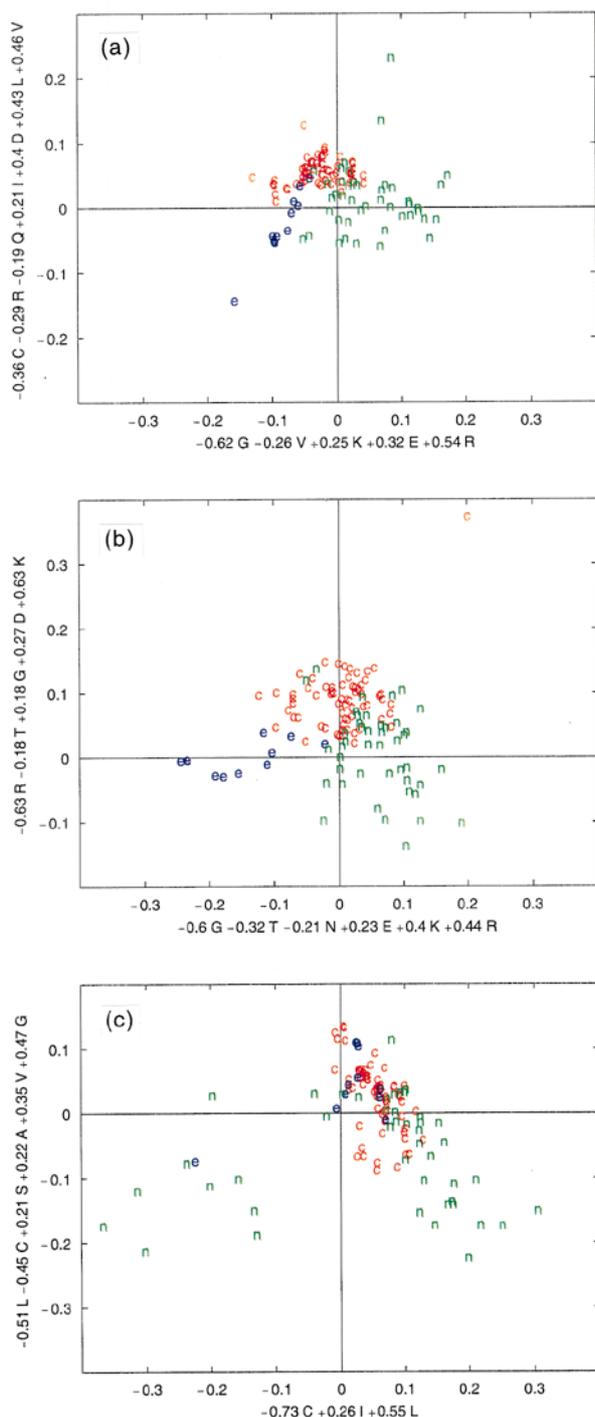
Figure 2. Principal component projections of the single data set. The figure shows the distributions of (a) total amino acid composition vectors, (b) surface composition vectors, and (c) interior composition vectors for the proteins in the *single* data set. The 20-dimensional composition vectors have been projected onto the plane described by the first two principal components (x-axis and y-axis respectively), where the principal components were calculated separately for each of the three vector sets. The axis labels indicate the amino acid types that contribute most significantly to the two principal components. The position of each vector is marked with a coloured letter, n (green), c (red), or e (blue), indicating the subcellular location of the protein (nuclear, cyto-

Note that the vectors defining the plane are fairly similar to those in the principal component projection of the surface composition of the *Single* data set (compare the axes labels for Figure 2(b) and Figure 3). The low degree of cluster overlap suggests that the projection can be used for predicting subcellular location based on the surface composition alone.

We tested this prediction method by projecting the surface composition vectors for the *Non-located* data set onto the plane in Figure 3 (see Supplementary material). Since this plane gives maximal separation between the class centres, and since the centres are calculated using many more data points than in the *Single* data set, this plane should be better for prediction than the plane in Figure 2(b). The proteins that occurred in the following regions were regarded as strongly predicted as belonging to the corresponding class: $x > 0$, $y < 0.01$ (nuclear); $x < 0.02$, $y > 0.08$ (cytoplasmic); $x < -0.06$, $y < 0.02$ (extracellular). Of the 262 vectors from the *Non-located* data set, 116 fell into the strongly predicted regions. By manually checking the database entries, we were able to assign 84 of the strongly predicted proteins to the three major subcellular locations. Of the assigned proteins, 65 (77%) were predicted to be in the correct location.

We also projected the surface composition vectors from the *Glycosylated* data set onto the same plane (see Supplementary material). All the glycosylated proteins are extracellular, but the vectors occurred in all three regions of the projection, mostly in the cytoplasmic and extracellular region. This result is consistent with the previously proposed hypothesis that glycosylation is a general mechanism to alter the surface properties of proteins that evolved initially in the cell interior, so that they are adapted to the extracellular environment (e.g. see Wagh & Bahl, 1981).

The difference between the surface composition of the three locations was mostly in amino acids with charged side-chains (see axes labels in Figure 3). Figure 4(a) shows the average surface composition vectors for each of the three location classes. In Figure 4(b), this information is summarized by grouping together amino acid types with similar electrostatic properties. For nuclear and cytoplasmic proteins, the total percentage of charged and polar surface residues is similar; cytoplasmic proteins have equal numbers of positively and negatively charged residues, whereas nuclear

plasmic, or extracellular, respectively). The total composition vectors show clustering by location class; this is even clearer in the surface composition vectors; however, the interior composition vectors show no tendency to cluster. The surface composition vectors for two nuclear proteins (around $x = -0.5$, $y = 0.1$ in *b*) fall well into the cytoplasmic cluster. The two structures used to calculate these data were of domains, not complete proteins, hence these outliers may be explained by missing interactions that in the complete protein would bury some of the residues exposed in the domain structure.
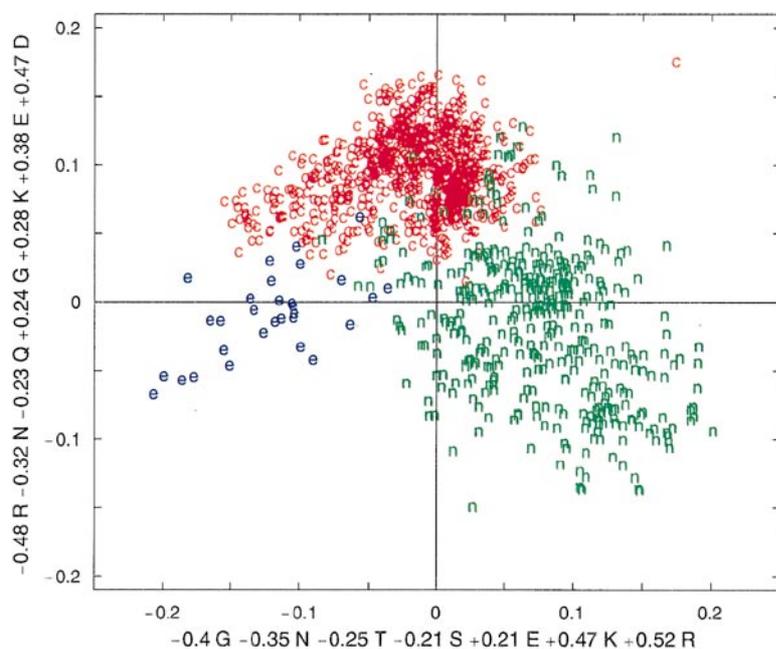
Figure 3. Surface composition vectors for the *homology* data set. The surface composition vectors for the *homology* data set are projected onto the plane defined by the three average surface composition vectors (one for each location class). Vector positions are marked with coloured letters to indicate the subcellular location (as in Figure 2). The axes are labelled in as Figure 2.

proteins have more positively charged residues. In marked contrast, extracellular proteins have significantly fewer charged surface residues, with about equal numbers of positively and negatively charged residues. The lack of charged residues is compensated by an increase in polar residues. Proteins in all three locations have almost the same proportion of apolar surface residues.

Histograms of solvent accessibility for each amino acid type differed between the three locations (see Supplementary material). The differences were correlated with the vectors defining the plane separating the clusters (Figure 3). Compared to proteins from the two other classes, nuclear proteins tended to have fewer completely buried residues, suggesting that nuclear proteins may be smaller on average. We confirmed this trend by comparing the distribution of protein lengths between the three locations (see Supplementary material). However, the trend was clearly not strong enough to be used alone for predicting location from sequence.

Discussion

While the subcellular locations of almost all the proteins in the PDB would be known, they have not been entered into either the PDB or SWISS-PROT. This has greatly limited the number of structures we could use for this study. For this reason, we were not able to exclude partial structures (structures of a domain rather than the whole protein). Hence some exposed residues may actually be buried in the complete structure, and the separation between the clusters may actually be better than indicated in Figure 3. However, the current data sets clearly established that the surface composition vectors had a strong signal indicating the subcellular location. Since we observed this signal in both the *Single* and *Homology* data sets, it is very unlikely to have arisen from bias in the data sets. The total amino acid composition had a weaker location signal, while the interior composition had little or no location signal.

These results are consistent with our hypothesis that protein structures have adapted to constraints on the physio-chemical environment of each subcellular location. A second class of evolutionary constraints may be imposed by the different functional roles of proteins in different subcellular locations. For example, many proteins in the nucleus bind DNA (95% of the nuclear proteins in the *Single* data set were DNA-binding), and hence would be subject to a constant pressure to conserve surface residues favourable for DNA binding. In either case, the results suggest that the protein surface has been the focus of evolution, in agreement with the results of Lichtarge *et al.* (1996).

The difference in electrostatic properties of the surface of proteins in the three locations (Figure 4(b)) can be summarized as follows. In all three aqueous environments, about one-third of the surface residues were apolar. Of the remaining residues, the breakdown between polar and charged residues depended on the total ionic strength of the environment: the nucleus and cytoplasm have about the same total ionic strength and the same proportion of charged surface residues. Outside the cell, where the total ionic strength is much lower, the proportion of charged surface residues is also lower; this is compensated by an increase in polar surface residues. These may be general principles by which proteins adapt to their physio-chemical environment. In the near future, we intend to test these principles by considering

more subcellular location classes, and a broader range of organisms (Eubacteria and Archaea).

Cytoplasmic proteins have a balance of acidic and basic surface residues, while extracellular proteins have a slight excess of acidic surface residues. However, nuclear proteins have a pronounced excess of basic surface residues. This is clearly related to the large negative charge on the DNA, but there are (at least) two explanations. Firstly, the majority of nuclear proteins considered were DNA-binding, and in many the DNA-binding site will have an excess of basic residues to facilitate binding to the DNA phosphate backbone. Secondly, the excess positive charge on nuclear proteins may be a result from selective pressure to neutralise the overall negative charge in the nucleus.

The location signal seen in the surface composition was often strong enough to predict location class; this is effectively predicting location from tertiary structure. The level of prediction accuracy obtained is impressively high, given that the method used was completely unoptimised; the method could clearly be improved by using machine-learning techniques such as neural nets, and by including additional data. This ability to predict subcellular location of proteins given the tertiary structure may be of practical use. Increasingly, new proteins and domains are being discovered that are known to be important, either from their correlation with some disease state, or from their association with better-characterised proteins, but for which the exact function or location is unknown.
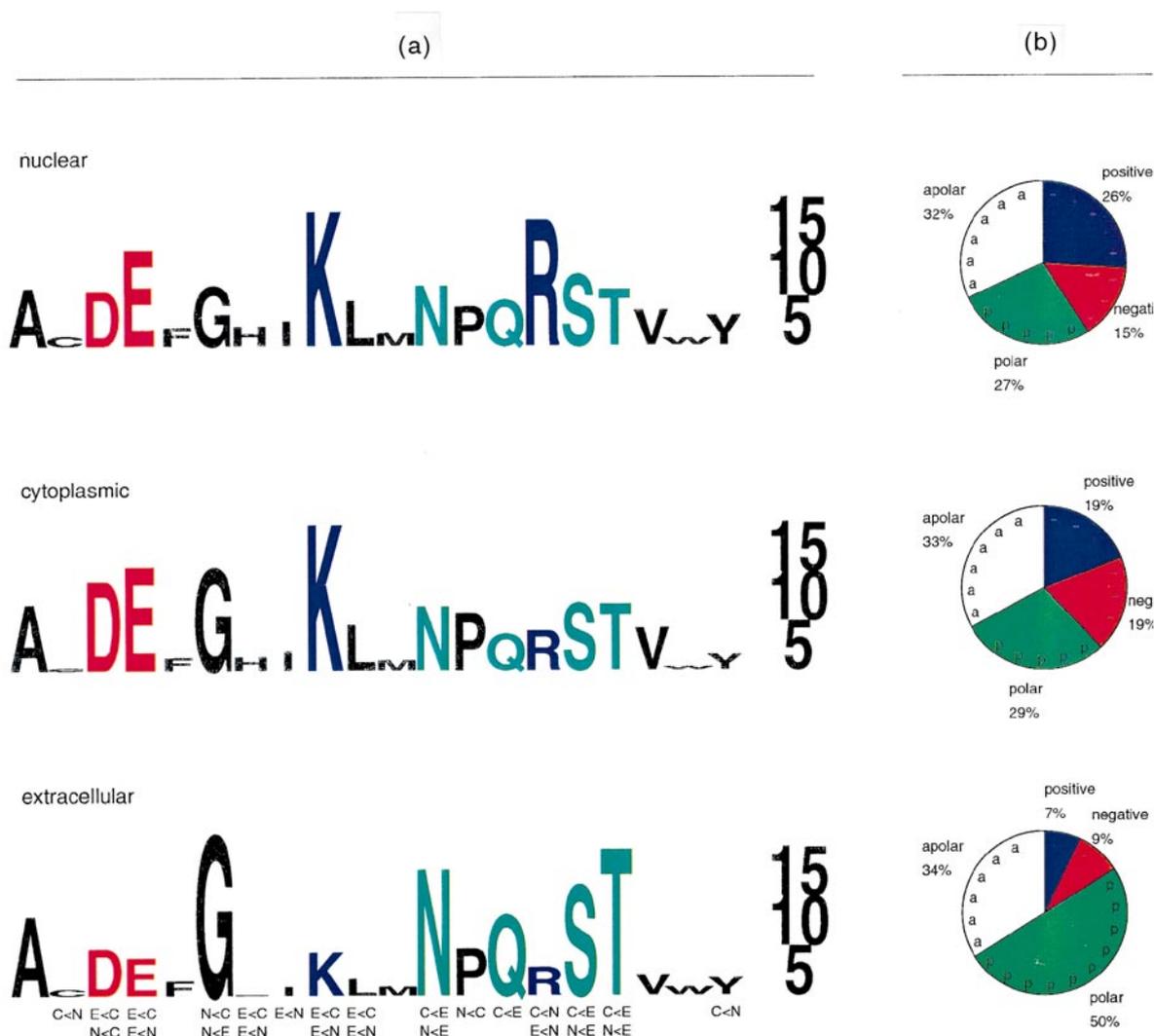

Figure 4. Average protein surface composition in different subcellular locations. (a) Components of the average surface composition vectors for nuclear (top), cytoplasmic (middle), and extracellular (bottom) proteins. The amino acids are represented by the single-letter code and are coloured by the electrostatic properties of their side-chains: black, apolar; red, acidic; blue, basic; and green, polar. The height of the letters is directly proportional to their contribution to the vector. The numbers in the right-hand column indicate the percentage contributions. The inequalities at the bottom of the Figure indicate which pairwise differences are statistically significant (using the mean difference test, 1% confidence level; e.g. under proline is the inequality $N < C$, indicating that the average composition of proline in the nucleus is significantly less than in the cytoplasm). (b) Pie chart representation of the same data as (a), except that amino acid types have been grouped by the electrostatic properties of their side-chains.

As a result, tertiary structures are being determined for proteins of unknown function and location (e.g. the PH domain: Macias *et al.*, 1994). In such cases, the strong correlation between surface composition and subcellular location may be useful as a step towards understanding a protein's function.

Our results also suggest that it may be possible to improve the prediction of location from sequence alone, using the clusters we observed in combination with a method for accurately predicting surface residues (Rost & Sander, 1994). Figure 2 suggests that this method would be more accurate than the existing methods that use total composition (Nishikawa *et al.*, 1983a,b; Cedano *et al.*, 1997). This method would also be complementary to methods based on signal peptides (Nakai *et al.*, 1988; Nakai & Karehisa, 1991, 1992) or methods based on homology. We intend to explore this idea in the future.

In conclusion, we have found a clear signal in the surface composition of protein structures that indicates the subcellular location; the signal was strong enough to allow accurate prediction of location. This supports our hypothesis that protein structures have adapted to the different physio-chemical environments in each subcellular location. The results suggest several principles that proteins use in adapting to particular physio-chemical environments; if these principles can be established by further studies, they may be useful in protein design. The results also suggest how the prediction of location from sequence may be improved.

Methods

Definition of the data sets

The protein structure database (PDB: Bernstein, 1977), and sequence databases such as SWISS-PROT (Bairoch & Apweiler, 1997) are highly biased towards particular protein families (Hobohm *et al.*, 1992). To reduce this bias, we selected our data sets in the following way. We started from the largest sequence-distinct subset of PDB (taken from the FSSP database: Holm & Sander, 1996); this subset is comprised of 849 protein chains chosen such that no pair has more than 25% pairwise sequence identity. From this subset we selected all globular eukaryotic proteins that occur in one of the three main subcellular locations (nucleus, cytoplasm, or extracellular space), based on the annotations in the SWISS-PROT entries. We excluded proteins from other subcellular locations (ribosome, mitochondria, chloroplast, vacuole, Golgi apparatus, endoplasmic reticulum, etc.), since our current purpose was to establish if a consistent difference could be observed for proteins from the three major location classes. In the remaining set of proteins, we distinguished between those that were and were not glycosylated, as glycosylation greatly affects the protein surface properties. Thus we defined two data sets: the *Glycosylated* data set (annotated in SWISS-PROT as glycosylated) and the *Single* data set (non-glycosylated, with a single member from each sequence family)

We then constructed an extended data set consisting of proteins sequences with known locations for which a structure could be modelled by homology. This *Homology* data set was constructed as follows: for each protein structure in the *Single* data set, we searched in SWISS-PROT for all eukaryotic protein sequences with ⩾ 40% pairwise sequence identity, and with known location (nuclear, cytoplasmic, extracellular). At this high level of sequence homology, it is safe to infer that these sequences have the same fold as the protein from the *Single* data set (Sander & Schneider, 1991); we chose such a high homology cut-off in order to guarantee conservation of solvent accessibility (Rost & Sander, 1994). Even at this high level of homology, it sometimes occurred that two members within the same sequence family had different subcellular locations; in such cases, the entire sequence family was excluded from the *Homology* data set. A final data set, the *Non-located* data set, was constructed with eukaryotic proteins in the sequence-distinct subset of PDB for which the subcellular location was not annotated in SWISS-PROT.

For the *Single*, *Glycosylated*, and *Non-located* data sets, the exposure state of each residue was calculated from the solvent-accessible surface area (Connolly, 1983) in the DSSP database (Kabsch & Sander, 1983). The surface area for each residue (in Å^2) was normalised by the maximal residue accessibility to yield a relative accessibility (as described by Rost & Sander, 1994). These values were then used to classify each residue as belonging either to the surface (relative accessibility ⩾ 25%) or the interior (relative accessibility < 25%) of the protein (Chothia, 1976; Hubbard & Blundell, 1987). For the *Homology* data set, the exposure state was calculated similarly except that the surface area values were inferred from the corresponding values in the homologous structure using the sequence alignments in the HSSP database (Sander & Schneider, 1994).

Analysis of composition vector

The total composition vector, $\mathbf{c}_i$, for a protein i is defined as the row vector $\mathbf{c}_i = \{c_{ij}\}$, where $j = 1, \ldots, 20$ indicates the amino acid type. The composition of the jth amino acid, c_{ij}, is defined as:

$$c_{ij} = r_{ij} \left/ \sum_{j=1}^{20} r_{ij} \right. \tag{1}$$

where R_{IJ} is the number of residues of amino acid type j in protein i. We also calculated surface and interior composition vectors, defined as above except that the R_{IJ} terms were then the number of residues of type j at the surface or in the interior, respectively.

For the *Single* data set, composition vectors were calculated for all proteins; these were then used to define a sample variance–co-variance matrix, **S**, as follows:

$$\mathbf{S} = \{s_{jk}\} = \left\{ \sum_{i=1}^{n} (c_{ij} - \bar{c}_j)(c_{ij} - \bar{c}_k)/n \right\} \qquad (2)$$

where:

$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} c_{ij} \qquad (3)$$

is the average composition of the *j*th amino acid type over the *n* proteins in the data set. The principal components of the set of composition vectors are then the Eigenvectors of **S** (e.g. see Anderberg, 1973). The composition vector for each protein was then projected onto the plane defined by the first two principal components using the standard inner product. This provides a two-dimensional view of how the component vectors are clustered (Figure 2). Note that in this analysis, the subcellular location class of the proteins is not considered. Hence, the resulting view will be unbiased in the sense that if the vectors are observed to cluster by location class, the clustering would be due to a trend in data, and not due to the analysis. Conversely, however, a cluster in the data may not be observed if the proteins belonging to this class are under-represented compared to the other classes.

For the *Homology* data set, we calculated a different projection that specifically aims to best separate the vectors from the three location classes into distinct clusters. This was done by projecting into the plane defined by the three average composition vectors (one for each location class). The plane was calculated as above except that only the three average vectors were used in defining **S**. Hence only two of the resulting 20 Eigenvectors had non-zero Eigenvalues. Then, the composition vectors for each protein in the *Homology* data set where projected into this plane; this plane was also used to project the composition vectors from the *Glycosylated* and *Non-located* data sets.

References

Anderberg, M. R. (1973). *Cluster Analysis for Applications,* Academic Press, New York.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucl. Acids Res.* **25**, 31–36.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Briggs, M. S. & Gierasch, L. M. (1986). Molecular mechanisms of protein secretion: The role of the signal sequence. *Adv. Protein Chem.* **38**, 109–180.

Cedano, J., Aloy, P., Pérez-Pons, J. A. & Querol, E. (1997). Relation between amino acid composition and celluar location of proteins. *J. Mol. Biol.* **266**, 594–600.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science,* **221**, 709–713.

Dingwall, C. (1991). Transport accross the nuclear envelope: enigmas and explanations. *BioEssays,* **13(5)**, 213–218.

Dingwall, C. & Laskey, R. A. (1986). Protein import into the cell nucleus. *Annu. Rev. Cell Biol.* **2**, 367–390.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–17.

Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucl. Acids Res.* **24**, 206–210.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159–171.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.

Macias, M. J., Musacchio, A., Ponstingl, H., Nilges, M., Saraste, M. & Oschkinat, H. (1994). Structure of the pleckstrin homology domain from β-spectrin. *Nature,* **369**, 675–677.

Nakai, K. & Kanehisa, M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct. Funct. Genet.* **11**, 95–110.

Nakai, K. & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics,* **14**, 897–911.

Nakai, K., Kidera, A. & Kanehisa, M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **2**, 93–100.

Nakashima, H. & Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **238**, 54–61.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.

Nishikawa, K., Kubota, Y. & Ooi, T. (1983a). Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem. (Tokyo),* **94**, 981–995.

Nishikawa, K., Kubota, Y. & Ooi, T. (1983b). Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J. Biochem. (Tokyo)*, **94**, 997–1007.

Pfeffer, S. R. & Rotheman, J. E. (1987). Biosynthetic protein transport and sorting by the endoplasmic reticulum and Golgi. *Annu. Rev. Biochem.* **56**, 829–852.

Rost, B. (1997). Protein structures sustain evolutionary drift. *Fold. Des.* **2**, S19–S24.

Rost, B. & O'Donoghue, S. I. (1997). Sisyphus and protein structure prediction. *Comput. Appl. Biosci.* **13**, 345–356.

Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20(3)**, 216–226.

Rost, B., Casadio, R. & Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **5**, 1704–1718.

Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Sander, C. & Schneider, R. (1994). The HSSP database of protein structure – sequence alignments. *Nucl. Acids Res.* **22**, 3597–3599.

Sjöström, M., Wold, S., Wieslander, Å. & Rilfors, L. (1987). Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. *EMBO J.* **6**, 823–831.

Verner, K. & Schatz, G. (1988). Protein translocation across membranes. *Science*, **241**, 1307–1313.

von Heijne, G. (1985). Signal sequences. The limits of variation. *J. Mol. Biol.* **184**, 99–105.

Wagh, P. V. & Bahl, O. P. (1981). Sugar residues on proteins. *CRC Crit. Rev. Biochem.* **10**, 303–377.