PROTEINS: Structure, Function, and Genetics 9:108–119 (1991)

The Frequency of Ion-Pair Substructures in Proteins Is Quantitatively Related to Electrostatic Potential: A Statistical Model for Nonbonded Interactions

Stephen H. Bryant and Charles E. Lawrence
Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201-0509

ABSTRACT **A statistical analysis of ion pairs in protein crystal structures shows that their abundance with respect to uncharged controls is accurately predicted by a Botlzmann-like function of electrostatic potential. It appears that the mechanisms of protein folding and/or evolution combine to produce a "thermal" distribution of local nonbonded interactions, as has been suggested by statistical–mechanical theories. Using this relationship, we develop a maximum likelihood methodology for estimation of apparent energetic parameters from the data base of known structures, and we derive electrostatic potential functions that lead to optimal agreement of observed and predicted ion-pair frequencies. These are similar to potentials of mean force derived from electrostatic theory, but departure from Coulombic behavior is less than has been suggested.**

Key words: protein structure, statistical analysis, ion pairs, electrostatic potential, maximum likelihood, maximum entropy

INTRODUCTION

The native structures of proteins are stabilized by many nonbonded interactions. Some local structures reflect the action of elementary forces, such as hydrogen bonding among acceptor and donor groups,[1] or electrostatic interaction among charged residues.[2–6] Others, such as the packing of hydrophobic residues in the protein interior,[7–9] appear to reflect aggregate interaction of the polypeptide with water.[10,11] The structural basis of conformer stability is understood at only a qualitative level, however. This is dramatically illustrated by computer experiments with misfolded proteins, which show that energy calculations using current atomic force fields cannot distinguish native conformers from grossly incorrect model structures.[12,13]

It has been found that misfolded proteins can be identified by calculations which measure the abundance of substructures common in native proteins, such as hydrophobic groups in contact with one another or polar groups accessible to water.[12–15] This suggests that conformer stability may to some extent be represented as a sum over the contributions of substructures, and that the data base of known structures contains information on what these are. It is unclear that there exist additive free-energy relationships as precise as in model systems,[10,11,16] however. It is also unclear how one should identify substructures for which this is a good approximation, or determine corresponding potentials from structural data. Favorable interactions occur more often than expected by chance,[3,4,6–9,17–22] but the probability model which relates substructure frequency to energy is unknown.[23]

Implications of an additive free-energy relationship for substructures have been considered before in the context of statistical–mechanical theories of polymer conformation.[20,24,25] In these theories conformer free energy is separated into additive contributions from backbone conformation and nonbonded interaction. The free energy of nonbonded interaction is expressed as a sum over the contributions of substructures involving close residue contacts, which are assumed to be additive and independent. For a random-sequence ensemble these assumptions imply that the relative statistical weights of substructures will be given by a Boltzmann factor in the difference in substructure chemical potential. A similar conclusion results from consideration of "selection entropy" arising from evolutionary sequence variation in biological macromolecules.[26] These theories may not be entirely applicable to proteins,[27–29] but they nonetheless suggest a simple hypothesis, that relative substructure frequencies should be predicted by a Boltzmann-like probability model.

In this paper we develop a statistical methodology that allows us to test whether observed substructure distributions are accurately described by this model. Using maximum likelihood theory, we find ener-

Received April 23, 1990; revision accepted July 11, 1990.
Address reprint requests to Dr. Stephen H. Bryant, Biometrics Laboratory, Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, NY 12201-0509.

getic parameters which lead to optimal agreement of the Boltzmann-like probability model with the data. We then use known statistical properties of maximum likelihood estimators to define confidence limits and to evaluate objectively the quantitative agreement of observed and predicted substructure frequencies. These methods also provide a means to test alternative probability models, and thus to derive empirical potentials which are most consistent with observed substructure frequencies.

We consider ion pairs[2–6] because these data afford a powerful test of the Boltzmann-like probability model. Electrostatic interaction in water is relatively well understood,[30–32] and may be represented by potentials of mean force which are functions of distance.[33–36] Furthermore, it is possible to identify control substructures that account for forces other than electrostatics, and to compare frequencies of substructures whose differences are large compared with the precision of crystallographic coordinates. We thus may attempt to predict the abundance of ion pairs relative to controls as a specific function of distance, and examine critically the agreement of the data and model. We also may compare the empirical potentials derived from these data to predictions from electrostatic theory and experiment.

METHODS

Statistical Model for Substructure Frequencies

The Boltzmann-like probability model suggested by statistical-mechanical theories[20,24–26] may be written

$$\frac{p(\hat{S}^i)}{p(\hat{S}^0)} = \exp[-\beta\delta\mu(\hat{S}^i, \hat{S}^0)] \qquad (1)$$

Here $p(\hat{S}^i)$ and $p(\hat{S}^0)$ are probabilities of observing substructures of types i and 0, and $\delta\mu(\hat{S}^i, \hat{S}^0)$ is the change in conformer free energy resulting from their exchange. β is a positive constant analogous to the term $1/kT$ in the Maxwell–Boltzmann equation. $\hat{S}^i$ and $\hat{S}^0$ refer in general to the chemical and positional parameters of two or more amino acid residues, where nonbonded interaction contributes significantly to conformer stability.

When i and 0 differ with respect to a subset of structural parameters (1) may be written

$$\frac{p(\hat{R}^i \mid \hat{X})}{p(\hat{R}^0 \mid \hat{X})} = \exp[-\beta\delta\mu(\hat{R}^i, \hat{R}^0 \mid \hat{X})] \qquad (2)$$

Here $\hat{R}^i$ and $\hat{R}^0$ are the structural parameters which distinguish i and 0, and $\hat{X}$ are structural parameters describing a protein environment where this substitution may occur. $\hat{R}^i$ and $\hat{R}^0$ may refer, for example, to the chemical type of a reference residue, and $\hat{X}$ to relative coordinates and types of neighbor residues.[4,6,19–21,25,26] It is useful to consider such "conservative" exchanges because they are known to occur in proteins, and hence may be described by (1). Furthermore, the dependence of potential difference on environmental parameters may be known from chemical physics or model systems. The energetic cost of substituting valine for threonine, for example, should depend on accessibility to water.[7–9,14]

The dependence of potential difference on assumed energetic parameters may be noted explicitly by writing (2) as

$$\frac{p(\hat{R}^i \mid \hat{X},\hat{\Theta})}{p(\hat{R}^0 \mid \hat{X},\hat{\Theta})} = \exp[-\beta\delta\mu(\hat{R}^i,\hat{R}^0 \mid \hat{X},\hat{\Theta})] \qquad (3)$$

$\hat{\Theta}$ may refer, for example, to contact energies that vary according to residue-pair type.[15,20,25] In this form the Boltzmann-like probability model states that relative substructure frequencies should depend on structural differences, $\hat{R}^i$, $\hat{R}^0$, the protein environment, $\hat{X}$, and the parameters, $\hat{\Theta}$, in an appropriate potential function.

This hypothesis has been used in statistical mechanics to derive properties of a conformer ensemble given assumptions concerning the energies of nonbonded interaction. We consider the reverse problem. Given observed substructure frequencies, we wish to derive apparent energetic parameters $\hat{\Theta}$ and test whether the structural data are consistent with a Boltzmann-like probability model. For this purpose it is useful to adopt a data-analytic strategy based maximum likelihood estimation.[37] This procedure finds best-fit parameter values, and in addition provides access to the likelihood ratio test statistic. This statistic may be used to evaluate goodness-of-fit, define confidence limits on $\hat{\Theta}$, and determine whether alternative potentials and/or substructure definitions lead to better agreement of observed and predicted substructure frequencies.

Given that a substructure of type $i = \{0,1, \ldots, I\}$ occurs at a particular site, it is easily shown from (3) that individual substructure probabilities may be written

$$p(\hat{R}^i \mid \hat{X},\hat{\Theta}) = \frac{\exp[-\beta\delta\mu(\hat{R}^i, \hat{R}^0 \mid \hat{X},\hat{\Theta})]}{1 + \sum_{i=1}^{i=I} \exp[-\beta\delta\mu(\hat{R}^i, \hat{R}^0 \mid \hat{X},\hat{\Theta})]} \qquad (4a)$$

$$p(\hat{R}^0 \mid \hat{X},\hat{\Theta}) = \frac{1}{1 + \sum_{i=1}^{i=I} \exp[-\beta\delta\mu(\hat{R}^i, \hat{R}^0 \mid \hat{X},\hat{\Theta})]} \qquad (4b)$$

These expressions constrain substructure probabilities to sum to 1. Type 0 is designated arbitrarily as a reference state, such that the potential difference corresponds to an exchange of 0 for i. Given N^i substructure observations of each type, the log-likelihood for a data set $S = \{\hat{R}^i_j, \hat{X}_j, i = 0,1, \ldots, I, j = 1,2, \ldots, N^i\}$ may now be written

$$LnL\,(\hat{\Theta} \mid S) = \sum_{i=0}^{i=I} \sum_{j=1}^{j=N^i} \ln\left[p\,(\hat{R}^i_j \mid \hat{X}_j, \hat{\Theta})\right] \qquad (5)$$

The values of $\hat{\Theta}$ which maximize this expression are the maximum likelihood estimates, $\hat{\Theta}^{max}$.

For a goodness-of-fit test data for each substructure type are separated into B bins according to the value of $\hat{X}_j$. Intervals should be chosen such that there are meaningful differences among observed substructure counts in each bin, $n^i_j, j = 1, \ldots, B$. Predicted counts $e^i_j, j = 1, \ldots, B$ are obtained by summing (4) over observations in each bin, using parameters $\hat{\Theta}^{max}$. The likelihood ratio test statistic may now be written

$$LRTS = 2\left[\sum_{i=0}^{i=I} \sum_{j=1}^{j=B} n^i_j \log\left(n^i_j / e^i_j\right)\right] \qquad (6)$$

This statistic is asymptotically chi-squared with $(I)(B-Q)$ degrees of freedom, where Q is the number of free parameters $\hat{\Theta} = \{\Theta_1, \ldots, \Theta_Q\}$ in the model used to obtain predicted counts.[38] The probability that differences in observed and predicted counts are statistically significant may thus be obtained from standard tables. It has been shown that (6) is asymptotically equivalent to a conventional X^2 goodness-of-fit statistic.[38]

The likelihood ratio statistic may be used more generally to test hypotheses represented by specific values of $\hat{\Theta}$. The test statistic for the hypothesis $\Theta_i = \Theta^0_i, i = 1,2, \ldots, P, P<Q$, may be written

$$LRTS = 2\left[\underset{\Theta_1, \ldots, \Theta_Q}{Max}\, LnL\,(\Theta_1, \ldots, \Theta_Q \mid S)\right] \qquad (7)$$

$$-2\left[\underset{\Theta_{P+1}, \ldots, \Theta_Q}{\text{Max}} LnL(\Theta^0{}_1, \ldots, \Theta^0{}_P, \Theta_{P+1}, \ldots, \Theta_Q \mid S)\right]$$

This statistic is known to be asymptotically chi-squared with $P-Q$ degrees of freedom.[37] A confidence interval for Θ_i may be defined by calculating the probability of $\Theta_i = \Theta^\circ_i$ for various values Θ°_i. Alternative potentials may be evaluated by devising a general model which reduces to alternatives in question for specific values of $\hat{\Theta}$. These may involve dependence on alternate environmental parameters $\hat{X}$, or an alternate functional dependence on $\hat{X}$. It may be shown that (6) is a special case of (7), with observed counts treated as a "model" that reproduces bin counts exactly.

Ion-Pair Data

Crystallographic data are from the Protein Data Bank.[39] We consider a set of 143 proteins selected on the basis of having distinct amino acid sequences and nearly complete atomic coordinates:

1155C	1156B	1ABP	1ACX	1AZU	1BP2	1CAC	1CC5	1CCR	1CI2	1CRN	1CSE
1CTF	1CTX	1CY3	1CYC	1ECD	1ETU	1FB4	1FBJ	1FC1	1FDH	1FDX	1FX1
1GCN	1GCR	1GD1	1GP1	1GPD	1HDS	1HIP	1HMG	1HMQ	1IG2	1INS	1LH4
1LZ1	1MBD	1MBS	1MCP	1MEV	1MLT	1OVO	1P2P	1PCY	1PFC	1PHH	1PP2
1PPT	1PYP	1REI	1RHD	1RNT	1SGT	1SN3	1SNI	1TGN	1TIM	1TON	1TRM
1UBQ	1XY1	2ABX	2ACT	2ALP	2APP	2APR	2AZA	2B5C	2BP2	2CAB	2CCY
2CDV	2CGA	2CPP	2CTS	2CYP	2EST	2GN5	2HFL	2HHB	2INS	2LHB	2LYM
2LZM	2MHB	2MHR	2MT2	2OVO	2PAB	2PKA	2PRK	2PTN	2RHE	2SGA	2SNS
2SOD	2SSI	2STV	2TAA	2TBV	2WRP	3ADK	3C2C	3CLN	3CNA	3CPV	3DFR
3EBX	3FAB	3FXC	3GAP	3GPD	3GRS	3ICB	3PGK	3PGM	3RP2	3RXN	3SGB
3TLN	3WGA	451C	4ADH	4APE	4ATC	4CYT	4DFR	4FDI	4FXN	4LDH	4RHV
4SBV	4TNC	5API	5CHA	5CPA	5LDH	5RSA	5RXN	6PTI	8CAT	9PAP	

Results are similar when the analysis is repeated with subsets of these proteins selected according to crystallographic resolution, R-value, method of refinement, or lack of sequence homology.

To characterize ion-pair substructures we define a local frame of reference about one residue of each pair and tabulate relative coordinates of the other.[4,19,21] We accumulate data for eight ion-pair categories, (Asp,Glu)–(Arg,Asp,Glu,Lys), and eight matched control categories, (Asn,Gln)–(Arg,Asp, Glu,Lys). Aspartate and glutamate are replaced in the controls by asparagine and glutamine, residues which are similar in size, shape, and polarity but which lack charge.

The local frame of reference is defined by the carboxyl group of Asp or Glu or the amide group of Asn or Gln. Interresidue distances refer to a point midway between the carboxyl oxygen coordinates of Asp and Glu, a point midway between the amide oxygen and nitrogen of Asn and Gln, Lys amino nitrogen coordinates, and a point midway between the terminal guanidinium nitrogen coordinates of Arg. The categories Asp-Glu and Glu-Asp are indistinguishable with respect to ion-pair distances, and to avoid double counting only Asp-Glu is considered, with data for Asn-Glu and Gln-Asp treated as control.

We consider substructures with interresidue distances between 4 and 10 Å, a total of 7329 ion-pair

and 6783 control observations. The lower distance limit of 4 Å group separation corresponds to van der Waals contact of individual atoms for most orientations, and there are few observations at shorter distances. Average closest approach distance is less for substructures involving lysine, where the charged group is less bulky, and the lower distance limit may be reduced to 3 Å with little effect on results.[40] The 10 Å upper limit is a point at which there appears to be no longer a great effect of electrostatic interaction.[3] It may be decreased to 7 Å with little effect.

Results are similar when the individual atoms of Asp and Glu carboxyl and Arg guanidinium groups are treated as charge centers, and when subsets from the residue pair data are selected according to separation in the amino acid sequence or atomic thermal factors. Modeling results are similar when C_α coordinates are considered in place of functional group coordinates, suggesting that the relative frequency distributions are largely determined by the sequence and backbone conformation of the proteins in the sample.

Survey of the structural data base has been carried out using the PKB program system,[40] operated on a Sun 386i computer.

Ion-Pair Statistical Analysis

We treat the frequency distribution of control substructures as a reference state, modeling the relative frequency of ion pairs as a function of difference in electrostatic potential. With this design we implicitly average over the microscopic environments of many residue pairs, and therefore assume that electrostatic interaction may be represented as a function of distance and charge only.[34,41] To check that the controls account for systematic effects of other forces we compare relative substructure frequencies among the individual ion-pair and control categories. Forces other than electrostatics should not act similarly across these categories, since their constituent residues differ in chemical properties other than charge, and these comparisons test whether the effects of other forces tend to cancel as expected.

We consider five potential functions intended to represent average charging work in the presence of water and/or protein.[33–36] We refer to these as potentials, but they correspond to potentials of mean force, that is to free energies.[42] They may be written

$$E(d,q_1,q_2,\epsilon,\sigma) = 332 q_1 q_2 / d\epsilon F(d,\sigma) \text{ kcal/mol} \quad (8a)$$

Here d is the distance to a neighbor residue with charge q_2, which we assume to be -1 or $+1$ electron equivalents for (Arg, Asp, Glu, Lys). q_1 is assumed to be -1, since (8a) is to represent potential difference for substitutions (Asn, Gln) to (Asp, Glu). Energetic parameters are ϵ, an effective dielectric constant, and σ, a secondary screening parameter. $F(d,\sigma)$ corresponds to the fractional change in effective dielectric constant with distance, the term which differs among the five potentials. For a Coulombic potential we have $F(d,\sigma) = 1$.[34] For the others

$$F(d,\sigma) = \epsilon(d,\sigma)/\epsilon(6.5,\sigma) \quad (8b)$$

The meaning of σ depends on the specific form of the distance-dependent dielectric function $\epsilon(d,\sigma)$, as described below. Normalization by the value of $\epsilon(d,\sigma)$ at $d = 6.5$ Å allows the overall dielectric constant ϵ to interpreted as an equivalent energy-scaling parameter for all models, and reduces to near zero its correlation with σ.

To check that the control substructures are a valid reference state we hypothesize that potential difference according to (8) should depend only on charge and not on residue chemical type. Accordingly, we treat residue-pair type (Asp/Asn, Glu/Gln)− (Arg, Asp, Glu, Lys) as a categorical variable describing the protein environment, and test whether models with category-specific values of ϵ and σ lead to significantly improved agreement of observed and predicted substructure frequencies. To simplify notation we omit category subscripts in (8) and discuss the equivalent hypothesis that pooling of substructure data for individual ion-pair and control categories is justified.

For this design the Boltzmann-like probability model (3) may be written

$$\frac{p\,(R^1|d,q_1,q_2,\beta,\epsilon,\sigma)}{p\,(R^0|d,q_1,q_2,\beta,\epsilon,\sigma)} = \exp\left[-\beta\ 332\ q_1 q_2/d\epsilon F(d,\sigma)\right] \quad (9)$$

Here $p(R^1|d,q_1q_2,\ \beta,\epsilon,\sigma)$ and $p(R^0|d,q_1,q_2,\ \beta,\epsilon,\sigma)$ are probabilities of observing an ion-pair or control substructure, that is the probability that the reference residue is (Asp, Glu) as opposed to (Asn, Gln). By collection of constant terms (9) may be rewritten

$$\frac{p\,(R^1|d,q,\theta,\sigma)}{p\,(R^0|d,q,\theta,\sigma)} = \exp\left[\theta q/dF(d,\sigma)\right] \quad (10)$$

Here q corresponds to q_2, and θ to the collection of terms $\beta\ 332/\epsilon$. The change of sign in the exponent follows from $q_1 = -1$. It is apparent from this expression that β may be estimated from the data only when ϵ is known, or vice versa. Rather than assign particular values, we estimate in the analysis the parameter θ, and consider its interpretation as the ratio $\beta\ 332/\epsilon$.

Following (4), ion-pair and control substructure probabilities are written

$$p(R^1 \mid d,q,\theta,\sigma) = \frac{\exp\left[\theta q/dF(d,\sigma)\right]}{1 + \exp\left[\theta q/dF(d,\sigma)\right]} \quad (11a)$$

$$p(R^0 \mid d,q,\theta,\sigma) = \frac{1}{1 + \exp\left[\theta q/dF(d,\sigma)\right]} \quad (11b)$$

Following (5), the log-likelihood for the data set of N^1 ion pairs and N^0 controls is written

$$LnL\,(\theta,\sigma \mid S) = \sum_{j=1}^{j=N^1} p(R^1 \mid d_j,q_j,\theta,\sigma) \quad (12)$$
$$+ \sum_{j=1}^{j=N^0} p(R^0 \mid d_j,q_j,\theta,\sigma)$$
$$= \sum_{j=1}^{j=N^1} \theta\, q_j/d_jF(d_j,\theta,\sigma) -$$
$$\sum_{j=1}^{j=N^1} \ln\{1 + \exp\left[\theta\, q_j/d_jF\,(d_j,\sigma)\right]\}$$
$$- \sum_{j=1}^{j=N^0} \ln\{1 + \exp\left[\theta\, q_j/d_jF\,(d_j,\sigma)\right]\}$$

Values θ^{max} and σ^{max} that maximize this expression are found numerically, using Fortran programs called from PKB.[40]

The predicted proportion of ion-pair and control substructures at any d,q may be obtained from (10) using maximum likelihood parameter estimates. To evaluate predictions, however, it is necessary to sum (11a) and (11b) over observations in some distance interval, to generate predicted counts which may be compared to those observed. For a graphical evaluation we carry out summations in the manner of a cumulative distribution function, taking intervals from the minimum d out to each d at which there is a substructure observation. For a numerical goodness-of-fit test we divide the data into six intervals beginning at 4, 5, 6, 7, 8, and 9 Å interresidue distance. We then calculate the likelihood ratio statistic (6), which is chi-squared with five degrees of freedom for the Coulombic model and four degrees of freedom for the others.

To define confidence limits for θ we write (7) as

$$LRTS = 2\left[\underset{\theta}{Max}\, LnL\,(\theta,\sigma^{max}|S) - LnL\,(\theta',\sigma^{max}|S)\right] \quad (13)$$

This statistic is chi-squared with one degree of freedom, and we obtain the probability of $\theta = \theta^o$ from standard tables. Confidence limits for σ are calculated similarly. Models which contain a secondary screening parameter σ are written in such a way that they reduce to the Coulombic model as σ approaches zero. To determine whether their improved agreement with the data is statistically significant we therefore write (7) as

$$LRTS = 2\left[\underset{\theta,\sigma}{Max}\, LnL\,(\theta,\sigma|S) - \underset{\theta}{Max}\, LnL\,(\theta,\sigma^0|S)\right] \quad (14)$$

The second term corresponds to the maximum log-likelihood for the Coulombic model. This statistic is chi-squared with one degree of freedom. To determine whether category-specific values of ϵ and σ lead to significantly improved agreement we write (7) as

$$LRTS = 2\left[\sum_{k=1}^{k=K} \underset{\epsilon_k,\sigma_k}{Max}\, LnL\,(\epsilon_k,\sigma_k|S_k) - \underset{\epsilon,\sigma}{Max}\, LnL\,(\epsilon,\sigma)|S\right] \quad (15)$$

Here subscripts k indicate residue-pair categories within $S = \{S_1,S_2,\ldots,S_k\}$, for example (Asp/Asn)–Lys or (Glu/Gln)–Arg. This statistic is chi-squared with $2(K-1)$ degrees of freedom, or $K-1$ degrees of freedom for the Coulombic model.

In the statistical analysis we do not apply an explicit correction for oversampling of larger distances as a function of volume,[4,21] since this correction is implicit in the comparison of ion pairs and controls. We also do not apply explicit corrections for overall residue abundance, since these tend to cancel in the analysis, and results for pooled data are not significantly affected. Cancellation of neighbor-residue abundance corrections is implicit in comparison of ion pairs and controls. Cancellation of reference-residue abundance corrections arises from pooling of residue-pair categories with opposite signs in the potential function, since opposite q in (10) is equivalent to inversion of the Asp/Asn or Glu/Gln ratio. The latter cancellation thus applies to pooled data but not to individual categories.

RESULTS

Sections from the relative coordinate data are shown in Figure 1. Data for individual residue-pair categories are pooled in the figure according to whether they are a (−)(+) ion pair, (0)(+) control, (−)(−) ion pair, or (0)(−) control. Substructure frequencies clearly appear to reflect electrostatic interaction. There are more (−)(+) ion pairs than (0)(+) controls at short distances, and fewer (−)(−) ion pairs than (0)(−) controls, the expected effects of favorable and unfavorable interaction. These differences diminish with distance as expected. Similar observations may be made for other sections and for the complete data viewed in three dimensions by molecular graphics. Aside from these differences the substructures identified in the survey appear diverse, and there is little indication that particular coordinates are uniquely preferred or excluded. It seems reasonable to model these differences in terms of a continuous probability function.

Figure 1 also illustrates how the choice of reference state is critical to the analysis. Neighbor-group densities for (0)(+) and (0)(−) controls are different, due presumably to differences in size, shape, and other properties of (+) and (−) neighbor groups. By comparison with these controls, data for (−)(+) and (−)(−) ion pairs suggest roughly opposite effects of electrostatic interaction, but this is not the case if they are compared with some other reference state, for example to uniform density, the reference state employed in treatments of ions in solution.[33] The controls appear to account for effects of forces other than electrostatics. The occluded volume effect of co-

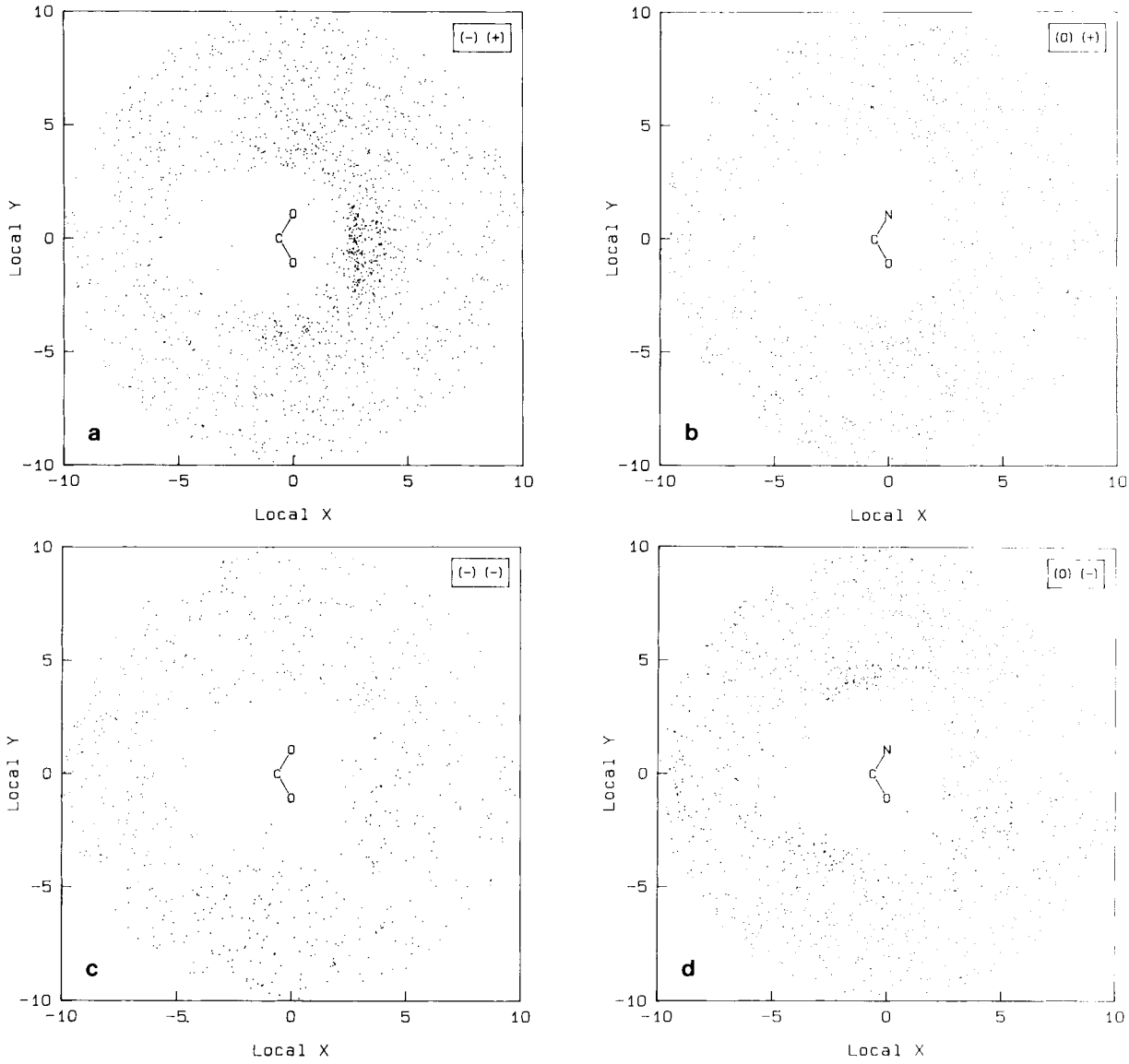

Fig. 1. Coordinates of charge-bearing neighbor groups relative to Asp/Glu carboxyl groups and Asn/Gln amide groups. Coordinates are in Å relative to an origin midway between the carboxyl oxygen atoms or amide oxygen and nitrogen atoms. The local *X* axis is defined by the bisector of the carboxyl carbon–oxygen bonds or amide carbon–oxygen and carbon–nitrogen bonds. *Z* is defined by the vector product of these bonds, and *Y* as the vector product of *Z* and *X*. The plots show sections from $Z = -2$ to $Z = 2$. Neighbor group coordinates refer to lysine amino groups, a point midway between the terminal nitrogen atoms of arginine guanidinium groups, and a point midway between the carboxyl oxygen atoms of aspartate and glutamate. Single-letter residue codes are Arg, R; Asn, N; Asp, D; Gln, Q; Glu, E; Lys, K. **(a)** shows (−)(+) ion pairs, *D–K, D–R, E–K, E–R*. **(b)** shows (0)(+) controls, *N–K, N–R, Q–K, Q–R*. **(c)** shows (−)(−) ion pairs, *D–D, D–E, E–E*. **(d)** shows (0)(−) controls, *N–D, N–E, Q–D, Q–E*.

valently bonded atoms in the reference-residue side chain is similar for ion pairs and controls. There are in each case fewer neighbor-group observations at small negative values of the local *X* coordinate. An analysis of spherical polar coordinates (not shown) indicates that angular distributions of neighbor groups are in general very similar, suggesting that the controls account for interactions with an angular dependence. It has been noted that hydrogen-bonding patterns of side-chain carboxyl and amide groups are similar.[1]

In Figure 2 relative frequencies of ion-pair and control substructures are plotted for a series of local distance intervals. The figure includes all observations, not only the sections in Figure 1, and data for individual residue-pair categories are shown separately. It may be seen that proportions of (−)(+) to (0)(+) observations are similar across categories and roughly the inverse of (−)(−) to (0)(−). This is again the expected effect of electrostatic interaction, further suggesting that comparison of ion pairs and controls isolates this effect. The plotting axes in Fig-

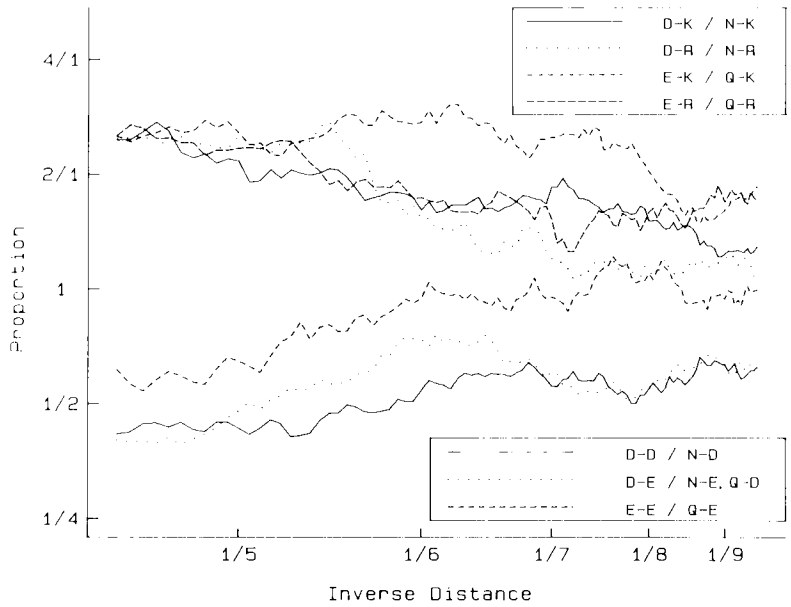

Fig. 2. Relative frequencies of charged neighbor groups as a function of distance from Asp/Glu carboxyl groups as compared to Asn/Gln amide groups. Substructure counts are taken for overlapping, 1 Å, distance intervals, centered on distances from 4.5 to 9.5 Å in steps of 0.05 Å. The inverse of the midpoints of these intervals are plotted on the horizontal axis. Relative substructure counts within each interval are plotted on a log scale on the vertical axis. Distances are taken between the origin and neighbor-group coordinate as described in the caption to Figure 1. Total numbers of substructure observations by category are *D–K*, 1504; *N–K*, 955; *D–R*, 1012; *N–R*, 768; *E–K*, 1413; *Q–K*, 664; *E–R*, 936; *Q–R*, 532; *D–D*, 771; *N–D*, 1375; *D–E*, 1115; *N–E*, 1059; *Q–D*, 812; *E–E*, 578; *Q–E*, 812. Overall residue abundances in the protein sample are *R*, 1351; *N*, 1906; *D*, 2194; *Q*, 1358; *E*, 1933; *K*, 2386.

ure 2 have been chosen to illustrate the approximately log-linear relationship of relative substructure frequency and inverse distance, the relationship predicted by (10) for a Coulombic potential.[34]

For quantitative modeling we employ the maximum likelihood procedure described above, considering probability models based on five dielectric screening functions:

$$\epsilon(d,\sigma) = 1, \quad \text{(M1)}$$
$$= \exp(\sigma d) \quad \text{(M2)}$$
$$= d^{\sigma} \quad \text{(M3)}$$
$$= 61 - 60\exp(-d/\sigma) \quad \text{(M4)}$$
$$= 78 - \{77(d/\sigma)^2\exp(d/\sigma)\,/\,[\exp(d/\sigma) - 1]^2\} \quad \text{(M5)}$$

Model 1 corresponds to a Coulombic potential.[34] Model 2 corresponds to Debye–Hückel theory.[33] The screening parameter σ is the inverse of the Debye length, which is 8 Å under physiological conditions. Model 3, with $\sigma = 1$, corresponds to an effective dielectric constant proportional to distance.[35] Model 4, with $\sigma = 10$ Å, corresponds to an empirical potential proposed by Warshel.[34] Model 5, with $\sigma = 2.5$ Å, corresponds to an empirical potential proposed by Hingerty et al.[36]

Modeling results for data pooled across residue-pair categories are shown in Figure 3. The plot displays observed and predicted proportions of ion-pair to control substructures computed at each successive distance observation, a cumulative proportion. The observed proportion is a good representation of the "curve" fitted by the maximum likelihood procedure, since each observation is weighted equally and data are not divided into arbitrary intervals. Distance is plotted on a scale that is linear in the number observations, so that agreement of observed and predicted proportions may be judged by comparison. Parameter estimates for each of the predicted curves are listed in Table I, together with $p > 0.95$ confidence limits obtained from likelihood ratio statistics (13).

It may be seen that the Coulombic model accounts for the general shape of the relative frequency curve. Predicted and observed proportions disagree by about 10%, however, for distances between 4 and 5 Å. With models 2 through 5 predicted and observed proportions clearly agree well, differing by no more than a few percent in any region of the curve. The likelihood ratio test for goodness-of-fit (6) accepts models 2 through 5 at $p > 0.38$, and marginally accepts model 1 at $p = 0.12$. Likelihood ratio tests (14) indicate that the improvement of models 2 through 5 over model 1 is statistically significant at $p > 0.05$. The latter tests are more powerful, since departures from Coulombic behavior are represented in the single parameter σ. These modeling results indicate that the data are described very well by Boltzmann-like probability models, and that

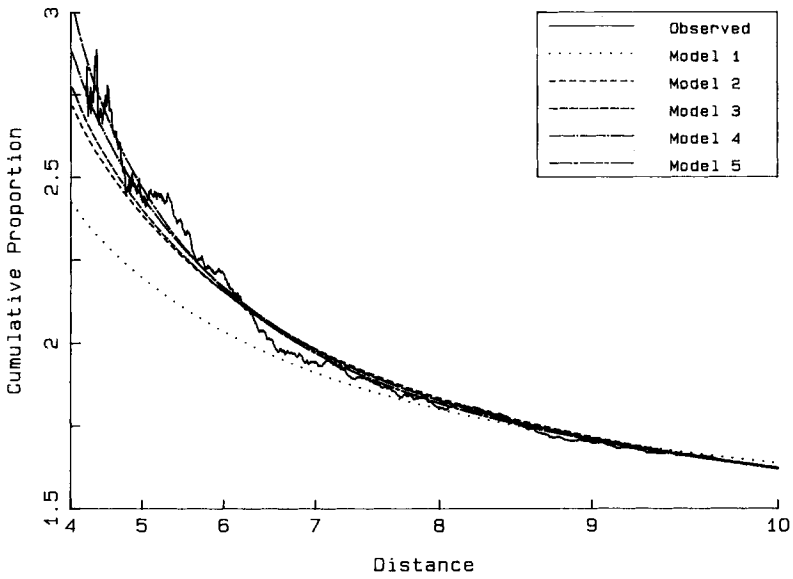

Fig. 3. Observed and predicted proportions of ion-pair and control substructures as a function of distance. $n(R^1|d\leq D)/n(R^0|d\leq D)$ is plotted versus D, where $n(R^1|d\leq D)$ is the observed or predicted count of ion pairs with interresidue distances less than or equal to D and $n(R^0|d\leq D)$ is the corresponding quantity for control substructures. This proportion is inverted for (−)(−) ion pairs and (0)(−) controls, with inversion of q in (11). With this manipulation proportions for all categories correspond to potential differences with the same sign, and they may be combined and displayed on a single plot. Curves for both observed and predicted counts are step functions, changing value at each observed d, but this is unapparent in the plot due to the large number of observations.

TABLE I. Maximum-Likelihood Parameter Estimates*

Model	Optimum θ	Optimum σ	Suggested σ
1	3.55 (3.31, 3.78)	—	—
2	3.58 (3.31, 3.85)	0.045 (0.002,0.091)	—
3	3.54 (3.27, 3.81)	0.31 (0.03,0.59)	1[35]
4	3.49 (3.23, 3.75)	3.2 (1.3,6.9)	10[34]
5	3.42 (3.16, 3.68)	0.95 (0.63,1.19)	2.5[36]

*Superscript numbers refer to references at the end of this article.

agreement is significantly improved by potential functions intended to represent screening effects.

It is difficult to judge quantitative agreement among individual ion-pair and control categories from Figure 2, but two or three categories are readily identified as outliers. Quantitative modeling confirms that these differences are statistically significant. Pooling of all seven categories is rejected at $p<0.05$ by the likelihood ratio test (15) using either model 1 or model 4. We use model 4 as a representative of the models containing a second parameter σ. These results suggest that mean electrostatic interaction may differ with residue-pair type, perhaps due to counterion binding, or alternatively that factors other than electrostatics may affect relative substructure frequencies among these categories.

The categories which differ most from the pooled data are Glu-Lys/Gln-Lys and Glu-Glu/Gln-Glu, which yield respectively larger and smaller estimates of the parameter θ. These differences indicate relatively fewer glutamine-containing substructures, and it seems likely that they reflect the lower overall abundance of glutamine as compared to glutamate. An abundance effect should cancel when these categories are subpooled, due to the opposite signs of the potential function, and we find that subpooling does lead to estimates of θ which agree with θ for pooled data.

The categories with the next largest differences are Asp-Arg/Asn-Arg and Asp-Asp/Asn-Asp, which yield smaller and larger estimates of θ, respectively. Examination of radial density traces suggests these differences are due to a small increase in the density of asparagine-containing substructures at distances between 7 and 9 Å (not shown). Consistent with this suggestion, we find that estimates of θ for both categories agree with θ for pooled data when only distances under 7 Å are included. Arg, Asn, and Asp are the residue types among those we consider where side chain hydrogen bonding is most often observed, and this distance range corresponds to networks involving intermediate groups.[3] It thus seems plausible to attribute these differences to a preference for asparagine in this position, perhaps

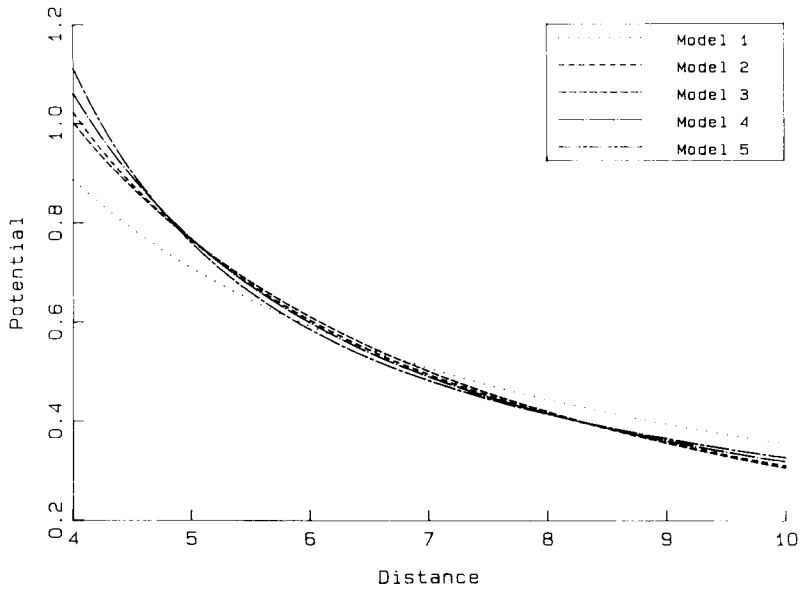

Fig. 4. Empirical ion-pair potential in kcal/mol as a function of distance. The functional forms of models 1 through 5 are described in the text. Maximum likelihood values of parameters θ and σ are from Table I. The energy scale of the potential depends on the value ϵ, which is here derived from the relationship $\epsilon = \beta\ 322/\theta$ with the assumption that β is 1 mol/kcal.

due to its role as both acceptor and donor. This effect should again cancel upon subpooling, and we find this is the case.

With these sub-poolings the likelihood ratio test (15) accepts pooling of the remaining five categories at $p>0.55$, indicating that remaining differences are below the level expected by chance. Given that significant differences may be associated with abundance and/or hydrogen bonding properties of (Asp,Glu) as opposed to (Asn,Gln), it seems most reasonable to conclude that they arise from these additional factors, rather than from differences in electrostatic interaction. Accepting this interpretation, we assume that data pooled across categories will provide the best description of electrostatic interaction, since sample size is largest and implicit correction for these effects is optimum.

Potential functions derived from maximum likelihood estimates of θ and σ for pooled data are shown in Figure 4. It may be seen that the improved agreement with the data of models 2 through 5 is due to the more rapid decrease of their potentials with distance. This is the effect these potentials were intended to represent,[33-36] which has its physical origin in the screening of charges by solvent, mobile ions, and/or other groups on the protein.[30-32] The substructure frequency data thus suggest corrections to a Coulombic potential which are qualitatively similar to those suggested independently from physical considerations.

The magnitude of the departure from Coulombic behavior is easily understood from model 2, the Debye–Hückel potential. The maximum likelihood estimate $\sigma = 0.045$ corresponds to a Debye length of 22 Å, which in turn corresponds to a 0.02 M NaCl solution at room temperature.[33] This is a small effect, in agreement with the proposal of Warshel[34] that departure from Coulombic behavior should be important only for short ion-pair distances. With maximum likelihood parameter values the potentials in models 2 through 5 are very similar to one another, and there is little indication that one functional form is to be preferred over another for this distance interval. The derived values of σ for models 3 through 5 do not agree quantitatively with those suggested by the authors of these models,[34-36] however, as shown in Table I. They correspond in each case to a less rapid decrease in electrostatic potential with distance.

The energy scale of the potentials in Figure 4 cannot be determined without an assumption concerning the effective dielectric constant ϵ or the Boltzmann-like parameter β. The parameter θ which we estimate corresponds to $\beta 332/\epsilon$, and one may solve for β only when ϵ is known, or vice versa. In Figure 4 we have assumed arbitrarily that β is 1 mol/kcal, and the energy scale on the vertical axis thus corresponds to $1/\beta$ kcal/mol/unit. In statistical-mechanical theories of polymer conformation[20,24,25] it has been assumed that $1/\beta$ is kT, 0.6 kcal/mol at 300° K. It is also reasonable to assume that the effective dielectric constant in the vicinity of charged groups is near that of water, since these groups fall almost entirely on the protein surface.[30-32]. With this as-

sumption $1/\beta$ for the Coulombic model is $332/(80\times3.55)=1.17$ kcal/mol. The physical bases of these assumptions are completely different, but either leads to potentials which are of the same order of magnitude as those suggested by detailed calculations and experiment.[30–32,41,43,44]

With the assumption that β is $1/kT$ one may derive an apparent dielectric constant for the Coulombic potential of $332/(0.6\times3.55)=156$. This corresponds to average pairwise interaction assuming unit charges, in the presence of other charged groups in solution and on the protein. The derived value is about twice the dielectric constant of water, in agreement with results from electrostatic theory and experiment which also suggest large effective screening constants.[30–32,41,43,44] We do not consider explicitly the microscopic factors responsible for dielectric screening, such as protonation or ion binding, and it is quite reasonable to believe that their effects should be apparent in the present analysis as a high effective dielectric constant.

These results from quantitative modeling support hypothesis (10). The relative frequencies of ion-pair and control substructures may be modeled well by a Boltzmann-like function of electrostatic potential, as shown in Figure 3. The empirical potentials derived from these data agree generally with those proposed on a physical basis in terms of their distance dependence and energy scale, as shown in Figure 4. Neither result proves that (10) is correct, but they are very suggestive when one considers that this hypothesis originates from statistical–mechanical theories which should apply approximately to proteins.[20,24–26]

DISCUSSION

It seems remarkable, given the complexity of protein structure, that the relative frequencies of ion-pair substructures appear to be related in a simple way to electrostatic potential. Boltzmann-like relationships have been suggested for other substructures,[7,9,17–20,22] and this may be a general property of proteins. Whatever the mechanisms responsible, this observation suggests that the data base of known structures provides not only examples of allowed substructures, but quantitative information on their contributions to conformer stability.

Statistical–mechanical theories[20,24–26] suggest that a Boltzmann-like relationship arises from the approximately additive contributions of substructures to conformer stability, and from the "equilibration" among substructures that occurs during protein folding and/or evolution. This is a simple and appealing suggestion as to the structural basis of protein stability and as to the manner in which this should be apparent from structural data. Additivity of substructure potentials is also suggested by the success of empirical energy functions in identifying misfolded proteins,[12–15] and by simulations which show that specific and cooperative folding reactions are predicted by this assumption.[45–48] Nonbonded interactions appear to be highly interdependent in proteins, however, due to their solid-like structures.[28] It is thus not obvious that these theories are applicable,[28,29] and it seems worthwhile to consider further the interpretation of the statistical results presented here.

Equation (10) is a maximum entropy probability model, which corresponds to the most probable distribution for energies constrained as to their average value, but otherwise subject to random variation.[42,49] The form of the distribution function thus suggests that partition of (Asp,Glu) and (Asn,Gln) among residue sites in the sample is systematically affected by electrostatic interaction but not by other factors. The positive values we infer for the parameter β indicate that this partition results in a net favorable interaction, consistent with observations concerning the importance of ion pairs in individual proteins[2] and with previous surveys which have shown favorable ion pairs to be more frequent than unfavorable ones.[3–6,15] The numerical value of β is determined only as the ratio $\theta = \beta 332/\epsilon$, but this nonetheless indicates substantial random variation. The choice of (Asp,Glu) over (Asn,Gln) is biased by only a factor of about 3 when another charged residue falls at a distance of 4 Å.

This is a "thermal" distribution in the sense that it is similar to that expected for free particles in a high-dielectric medium at room temperature. This similarity is apparent from the form of the distribution function and from the reasonable value for an effective dielectric constant which follows from the assumption that $1/\beta$ is 0.6 kcal/mol. The substructure data are not derived from a thermal system, however.[23] Crystallographic coordinates are time averages that provide no information on thermal fluctuations which proteins undergo. It has been proposed that the structures in the data base may be representative of a conformer ensemble,[48,50] but it is well known that proteins fold as a cooperative unit,[11] and thus that substructure distributions are in principle affected by global constraints on protein conformation, and not simply by available thermal energy.[23] It seems reasonable to assume that β reflects a relationship of energy and entropy, as in thermodynamics,[26,42,49] but with the understanding that statistical entropy arises from the various processes governing protein structures and from the techniques by which they are analyzed.

In statistical-mechanical treatments of polymers[20,24,25] the conformer distribution function is factored into terms which depend on backbone torsional parameters and terms which depend on nonbonded interactions, the "monomer gas."[24] The statistical weights of "monomer gas" configurations are expressed as a product of Boltzmann factors for substructures, under the assumption that independence

and additivity are good approximations if the strongest local interactions are considered, and potentials account for solvent effects. In a random-sequence conformer ensemble the relative statistical weights of substructures will be given by a Boltzmann factor in potential difference, since the other terms will cancel. These theories thus suggest that conformer equilibration should result in a "thermal" distribution for substructures, with β equal to $1/kT$, and the good agreement of the ion-pair data with this prediction suggests that they may apply to protein folding. One may argue that the observed distribution is too "hot", however, since the assumption $\epsilon = 80$ leads to $T = 332/(0.002 \times 80 \times 3.55) = 585°$ K. These theories most clearly apply to "molten globules,"[51,52] and it may be that their collapse into densely packed structures is apparent in an analysis of pairwise interactions as an additional entropy source.

It has also been suggested that a Boltzmann-like distribution for substructures may be a consequence of macromolecular evolution.[3,26] The basis for this suggestion is the observation that protein three-dimensional structure is conserved to a greater extent than amino acid sequence.[53–55] Substructure data may thus reflect primarily residue substitution events, which have occurred in the context of relatively fixed backbone structures. To model this evolutionary process one may again assume that conformer stability derives from independent contributions of substructures, which are randomly exchanged by mutation, but whose sum is constrained by selection. These are analogous to conditions in the classical derivation of the Maxwell–Boltzmann equation for a thermal system,[42] and they predict that relative substructure frequencies will follow a maximum entropy distribution. This suggestion is a generalization to protein tertiary structure of a selection theory developed by Berg and Von Hipple,[26] and the ion-pair modeling results are clearly consistent with its prediction. In this theory β is a measure of "selection entropy,"[26] analogous but not equivalent to $1/kT$. This may not be an important distinction, however. Native conformers are known to be only marginally stable,[11] and selection may not have "cooled" substructure distributions very much as compared to a random-sequence ensemble.[48,50]

There are also sources of "noise" that are intrinsic to the analysis of structural data. The crystallographic coordinates we consider have limited precision, for example, and this must contribute to the apparent statistical entropy. Our potential calculations ignore three-body and higher order interactions with the protein environment, and thus only approximate the energetic cost of substituting (Asp,Glu) for (Asn,Gln) in each substructure. The similarity of results for residue-pair categories and the good fit of the models and data indicate that higher order interactions have no systematic effect on ion-pair potential, but they may nonetheless contribute entropy and increase the apparent "temperature." The statistical entropy apparent in the data is about twice that predicted by $\epsilon = 80$ and $\beta = 1/kT$, $T = 300°$ K, and this may arise from imprecision in the coordinates and potential calculation. If ϵ is really near 160, which is not unreasonable, then the entropy in the data is near the "baseline" suggested by statistical–mechanical theories, and one may conclude that these sources of "noise" are relatively unimportant.

From these considerations it seems most reasonable to interpret the observed Boltzmann-like distribution as an indication that substructure "equilibration" has indeed occurred, as suggested by statistical–mechanical theories, but with the understanding that cooperative interactions may cause substructure distributions to appear somewhat "hotter" than these theories suggest. These modeling results certainly support the assumption that substructure contributions to conformer stability are approximately additive, since it is otherwise difficult to see why there should be any quantitative relationship at all between substructure frequency and energy.

An interesting implication of a Boltzmann-like probability model is that it should be possible to determine substructure potentials empirically, by statistical analysis of known protein structures. As a test case we have considered a simple pairwise interaction, where controls and functional forms for the potential were easily identified. The maximum likelihood techniques developed here are applicable to more complex substructures, however, and to refinement of more detailed and complete potentials. Such analyses will test the generality of the Boltzmann-like probability model.

An additive free energy relationship for protein substructures also has interesting implications for structure prediction. It may be possible to compute approximate conformer stabilities rapidly as a sum over local side chain interactions, that is, by measuring the "temperature" of their "monomer gas."[24] It has been shown, for example, that counts of implied hydrophobic contacts can distinguish correct from incorrect "alignments" of an amino acid sequence onto a hypothesized backbone fold.[15] It may be that potentials and "alignment" techniques can be refined to a point where plausible tertiary-structural models may be identified by scans of a conformer library.

REFERENCES

1. Baker, E.N., Hubbard, R.E. Hydrogen bonding in globular proteins. Prog. Biophys. Molec. Biol. 44:97–179, 1984.
2. Perutz, M.F. Electrostatic effects in proteins. Science 201: 1187–1191, 1979.
3. Wada, A., Nakamura, H. Nature of the charge distribution in proteins. Nature (London) 293:757–758, 1981.
4. Barlow, D.J., Thornton, J.M. Ion-pairs in proteins. J. Mol. Biol. 168:867–885, 1983.

5. Barlow, D.J., Thornton, J.M. The distribution of charged groups in proteins. Biopolymers 25:1717–1733, 1986.
6. Sundaralingam, M., Sekharudu, Y.C., Yathindra, N., Ravichandran, V. Ion pairs in alpha helices. Proteins 2:64–71, 1987.
7. Janin, J. Surface and inside volumes in globular proteins. Nature (London) 277:491–492, 1979.
8. Rose, G.D., Gaselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
9. Lawrence, C., Auger, I., Mannella, C. Distribution of accessible surfaces of amino acids in globular proteins. Proteins 2:153–161, 1987.
10. Baldwin, R.L. Temperature dependence of the hydrophobic interaction in protein folding. Proc. Natl. Acad. Sci. U.S.A. 83:8069–8072, 1986.
11. Privalov, P.L., Gill, S.J. Stability of protein structure and hydrophobic interactions. Adv. Protein Chem. 39:191–234, 1988.
12. Novotny, J., Bruccoleri, R.E., Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
13. Novotny, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
14. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature (London) 319:199–203, 1986.
15. Bryant, S.H., Amzel, L.M. Correctly folded proteins make twice as many hydrophobic contacts. Int. J. Peptide Protein Res. 29:46–52, 1987.
16. Gill, S.J., Wadsö, I. An equation of state describing hydrophobic interactions. Proc. Natl. Acad. Sci. U.S.A. 73:2955–2958, 1976.
17. Pohl, F.M. Empirical protein energy maps. Nature New Biol. 234:277–279, 1971.
18. Kolaskar, A.S., Prashanth, D. Empirical torsional potential functions from protein structure data. Phi- and psi potentials for non-glycyl amino acid residues. Int. J. Peptide Protein Res. 14:88–98, 1979.
19. Thomas, K.A., Smith, G.M., Thomas, T.B., Feldmann, R.J. Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments. Proc. Natl. Acad. Sci. U.S.A. 79: 4843–4847, 1982.
20. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
21. Burley, S.K., Petsko, G.A. Weakly polar interactions in proteins. Adv. Protein Chem. 39:125–189, 1988.
22. Janin, J., Miller, S., Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204: 155–164, 1988.
23. Bürgi, H.B., Dunitz, J.D. Can statistical analysis of structural parameters from different crystal environments lead to quantitative energy relationships? Acta Crystallogr. B44:445–448, 1988.
24. Lifshitz, I.M., Grosberg, A.Yu., Khokhlov, A.R. Some problems of the statistical physics of polymer chains with volume interaction. Rev. Mod. Phys. 50:683–713, 1978.
25. Dill, K.A. Theory for folding and stability of globular proteins. Biochemistry 24:1501–1509, 1985.
26. Berg, O.G., von Hipple, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. 193:723–750, 1987.
27. Flory, P.J. "Statistical Mechanics of Chain Molecules." New York: Interscience Publishers. 1969: 249–306.
28. Richards, F.M. Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
29. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
30. Warshel, A., Russell, S.T. Calculations of electrostatic interactions in biological systems and solutions. Q. Rev. Biophys. 17:283–422, 1984.
31. Honig, B.H. Hubbell, W.L., Flewelling, R.F. Electrostatic interactions in membranes and proteins. Annu. Rev. Biophys. Chem. 15:163–193, 1986.
32. Harvey, S.C. Treatment of electrostatic effects in macromolecular modeling. Proteins 5:78–91, 1989.
33. Bockris, J.O'M., Reddy, A.K.N. "Modern Electrochemistry," Vol. 1. New York: Plenum Press, 1970.
34. Warshel, A., Russell, S.T., Churg, A.K. Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability. Proc. Natl. Acad. Sci. U.S.A. 81:4785–4789, 1984.
35. Pickersgill, R.W. A rapid method of calculating charge-charge interaction energies in proteins. Protein Engineer. 2:247–248, 1988.
36. Hingerty, B.E., Ritchie, R.H., Ferrell, T.L., Turner, J.E. Dielectric effects in biopolymers: The theory of ionic saturation revisited. Biopolymers 24:427–439, 1985.
37. Kendall, M., Stuart, A. "The Advanced Theory of Statistics. Vol. 2. Inference and Relationship." New York: Macmillan, 1979.
38. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. "Discrete Multivariate Analysis: Theory and Practice." Cambridge, MA: MIT Press. 1975.
39. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. Protein data bank. In: "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. eds. Bonn, Chester, Cambridge: Int. Union of Crystallography, 1987: 107–132.
40. Bryant, S.H. PKB: A program system and data base for analysis of protein structure. Proteins, 5:233–247, 1989.
41. Gilson, M.K., Honig, B.K. Energetics of charge-charge interactions in proteins. Proteins 3:32–52, 1988.
42. Hill, T.L. "An Introduction to Statistical Thermodynamics." Reading, MA: Addison-Wesley, 1960.
43. Bashford, D., Karplus, M., Canters, G.W. Electrostatic effects of charge perturbations introduced by metal oxidation in proteins. A theoretical analysis. J. Mol. Biol. 203: 507–510, 1988.
44. Sternberg, M.J.E., Hayes, R.F., Russell, A.J., Thomas, P.G., Fersht, A.R. Prediction of electrostatic effects of engineering of protein charges. Nature (London) 330:86–88, 1987.
45. Krigbaum, W.R., Komoriya, A. Local interactions as a determinant for protein molecules: III. Biochem. Biophys. Acta 576:229–246, 1979.
46. Taketomi, H., Kano, F., Go, N. The effect of amino acid substitution on protein-folding and -unfolding transition studied by computer simulation. Biopolymers 27:527–559, 1988.
47. Skolnick, J., Kolinski, A., Yaris, R. Monte carlo simulations of the folding of β-barrel globular proteins. Proc. Natl. Acad. Sci. U.S.A. 85:5057–5061, 1988.
48. Lau, K.F., Dill, K.A. Theory for protein mutability and biogenesis. Proc. Natl. Acad. Sci. U.S.A. 87:638–642, 1990.
49. Hobson, A. "Concepts in Statistical Mechanics." New York: Gordon and Breach Science Publishers, 1971.
50. Ptitsyn, O.B. Protein as an "edited" statistical copolymer? In: "Conformation in Biology." Srinivasan, R., Sarma, R.H. eds. Guilderland, NY: Adenine Press, 1983: 49–58.
51. Shakhnovich, E.I., Finkelstein, A.V. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. Biopolymers 28:1667–1680, 1989.
52. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103, 1989.
53. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold. Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
54. Godzik, A., Sander, C. Conservation of residue interactions in a family of Ca-binding proteins. Protein Engineer. 2:587–596, 1989.
55. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. Nature (London) 339:31–36, 1989.