BLEEP—Potential of Mean Force Describing Protein–Ligand Interactions: I. Generating Potential

JOHN B. O. MITCHELL,[1] ROMAN A. LASKOWSKI,[2] ALEXANDER ALEX,[3] JANET M. THORNTON[1,2]

[1] *Department of Biochemistry and Molecular Biology, University College London, London, UK*
[2] *Department of Crystallography, Birkbeck College, University of London, London, UK*
[3] *Computational Chemistry, Pfizer Central Research, Sandwich, Kent, UK*

Received 18 January 1999; accepted 29 March 1999

ABSTRACT: We have developed BLEEP (biomolecular ligand energy evaluation protocol), an atomic level potential of mean force (PMF) describing protein–ligand interactions. The pair potentials for BLEEP have been derived from high-resolution X-ray structures of protein–ligand complexes in the Brookhaven Protein Data Bank (PDB), with a careful treatment of homology. The use of a broad variety of protein–ligand structures in the derivation phase gives BLEEP more general applicability than previous potentials, which have been based on limited classes of complexes, and thus represents a significant step forward. We calculate the distance distributions in protein–ligand interactions for all 820 possible pairs that can be chosen from our set of 40 different atom types, including polar hydrogen. We then use a reverse Boltzmann methodology to convert these into energy-like pair potential functions. Two versions of BLEEP are calculated, one including and one excluding interactions between protein and water. The pair potentials are found to have the expected forms; polar and hydrogen bonding interactions show minima at short range, around 3.0 Å, whereas a typical hydrophobic interaction is repulsive at this distance, with values above 4.0 Å being preferred. J Comput Chem 20: 1165–1176, 1999

Keywords: potential of mean force; knowledge-based potential; atomistic representation; protein–ligand interactions; computer-aided drug design

Present address: J. B. O. Mitchell, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK; e-mail: mitchell@biochemistry.ucl.ac.uk

Correspondence to: J. M. Thornton; e-mail: thornton@biochemistry.ucl.ac.uk

Contract/grant sponsor: Pfizer, Ltd.

CCC 0192-8651 / 99 / 111165-12

Introduction

Knowledge-based potentials of mean force (PMFs) have attracted a great deal of interest in recent years. Deriving these potentials involves converting distributions of interparticle distances from experimental systems into pair potential functions by assuming Boltzmann-like energetics.[2] This approach derives energy functions from experimental structural data, and contrasts with the alternatives of deriving potential energy functions from quantum chemistry,[3] or using parametric methods.[4] Sippl's seminal study[5] set out the methodology of the PMF approach, which leads to functions formally corresponding to the Helmholtz free energy, ΔA. He subsequently used these statistical potentials to describe the energetics of peptide hydrogen bonding[6] and atom pair interaction in proteins.[7]

Such potentials have found a number of uses in the study of proteins. Jones et al.[8] utilized them in threading, as did Reva et al.[9] O'Donoghue et al.[10] used PMFs as the basis of an approach aimed at the problem of general three-dimensional protein structure prediction. A review of the roles of empirical potentials in protein folding studies was presented by Vajda et al.[11] Also, Sippl[12] and Melo and Feytmans[13] utilized knowledge-based potentials in methods for detecting errors in protein structures, whereas Bahar and Jernigan[14] investigated the respective roles of hydrophilic and hydrophobic interactions in protein stability.

There has also been some controversy over the meaning, validity, and accuracy of database-derived statistical potentials. Thomas and Dill[15] suggested that the results obtained with knowledge-based potentials reflect little more than the burial of nonpolar groups and that such potentials would therefore be of limited value. Sunyaev et al.[16] concluded that the knowledge-based potentials used in threading do not discriminate adequately between the 20 amino acid types. Zhang and Skolnick[17] presented a more optimistic picture, believing that there is a meaningful relationship between derived and ''true'' potentials, and that statistical potentials can accurately predict residue properties such as secondary structure propensities.

Whereas several investigators have used empirical potentials in studies of protein–ligand binding, with Vajda et al.[11] reviewing some significant examples, very few of these studies have used knowledge-based PMFs. Verkhivker and coworkers[18,19] examined the binding of possible HIV-1 proteinase inhibitors using a knowledge-based PMF based only on HIV proteinase–inhibitor complexes. They also carried out a similar study on ligand complexes with the protein FKBP12.[20] Wallqvist et al.[21] carried out another study of HIV-1 proteinase inhibitor binding, using a function based on atom–atom surface propensities from a sample of 38 protein–ligand complexes. De Witte and coworkers[22,23] achieved some success with SMoG, a ligand design program incorporating a knowledge-based potential. Their potential function was, however, extremely coarse-grained, with atomic contacts being either on or off in a binary sense, depending on whether the atoms were within 5.0 Å.

We believe that the number of high-resolution protein–ligand complexes in the PDB is now sufficient to derive a PMF of general applicability in studies of protein–ligand binding. We develop such a potential here and utilize it in an accompanying study.[24] It is designed to facilitate the drug design process, in particular hit screening and lead optimization. We also intend it to be of wider use in more general studies of protein–ligand interactions. Whereas many PMFs, particularly those used in threading, have only one or two interaction sites per residue, ligand binding requires a finer level of detail. Thus, we opt for an atomic-level representation, similar to that used by Melo and Feytmans,[25] but our potential is for interactions with ligands as opposed to those within proteins. We cannot use a protein–protein potential for our purposes, because, first, many of the required atom types are not defined therein and, second, the natures of protein–protein and protein–ligand interactions are significantly different, even for equivalent atom types. The typical environment for a ligand is rather different from the protein core environment that dominates protein-derived PMFs. This is partly because the protein prefolds and has to be stable in solution, and also because the protein–ligand–solvent interface is a complex environment, quite different from the hydrophobic core of a protein. Our method, BLEEP (biomolecular ligand energy evaluation protocol) will use only data from protein–ligand interactions (the incorporation of water interactions is discussed in what follows), so that it is appropriate for calculating protein–ligand binding energies. Because the hydrophobic effect is known to be critical to many protein–ligand interactions, we expect that a PMF —implicitly including such entropic contributions

—should be able to outperform those theoretical methods that exclude entropy. BLEEP will be of much more general applicability than previous protein–ligand PMFs[19,20] that were derived from a very limited range of systems, and will also benefit from a very much more sensitive definition of atom types. Our pair potentials will be varying functions of interatomic distance, as opposed to the simple on–off model of DeWitte and Shakhnovich.[22]

Generation of Dataset

BLEEP is based on data from protein–ligand complex structures in the Brookhaven Protein Databank (PDB).[1] We restrict ourselves to high-resolution X-ray structures with resolutions of 2.0 Å or better; we do not include any NMR structures. All our structures contain noncovalently bound polyatomic ligand molecules.

To filter out homologous interactions, we represent the dataset in terms of {protein domain–ligand type} pairs. The definitions of domains were taken from the CATH classification of Orengo et al.[25] A ligand type covers all the chemically identical ligand molecules occurring in a given complex. The ligands varied in size from small species such as cyanide, sulfate, and phosphate to sizeable peptides and polysaccharides. The sample contains a good proportion of nucleotides and related molecules, as well as drug candidates and general organics. Overall, we find that this sample of ''biological'' molecules differs from a pseudorandom sample of organic molecules in having a significantly higher proportion of polar (noncarbon) heavy atoms; that is, 41% vs. 19%. We are publishing a detailed analysis of the chemical composition of this dataset elsewhere[27]; the details of the ligand molecules are also available over the internet.[28] Based on the resolution criteria, the dataset initially consisted of 2551 {protein domain–ligand type} pairs from 1117 PDB entries, prior to homology filtering. Two domains were considered homologous if they had identical CATH numbers at the first four levels (up to and including ''Homology''). No rigorous algorithm exists for determining an homology-like quantity for ligands, so a subjective evaluation of ligand similarity had to be made, aided by the structural diagrams provided in Laskowski et al.'s PDBSUM database.[29]

The 2551 pairs were grouped together by homology and, in all instances of homologous domains interacting with similar ligands, only the highest resolution representative was retained. The filtered dataset contained 594 {protein domain–ligand type} pairs from 351 PDB entries.[28] For practical purposes, we felt it better to incorporate whole chains in the dataset, so the whole of each of the chains containing these domains was included in the final dataset of 440 {protein chain–ligand type} pairs from 351 PDB entries (Table I). This protocol and the potentials so generated are collectively known as BLEEP-1.

As explained in what follows, a second set of potentials, this time including many more water interactions, and known as BLEEP-2, was calculated. The dataset for BLEEP-2 comprised 267 {protein domain–ligand type} pairs, corresponding to 222 {protein chain–ligand type} pairs from 188 PDB entries (Table II).[28]

There was only a small quantity of data for metals in our dataset. Thus, the use of a dataset with less restrictive homology criteria was necessary to enable potentials describing the interactions of metals to be calculated at all. This consisted of 198 proteins[28] having metal ions in their crystal structures.

Atom Types

There are two counterbalancing pressures that act in the defining of atom types. On the one hand, we would like to describe fully the chemical diversity of the dataset, and hence to define as many atom types as possible. On the other hand, we need as many data as possible to characterize the interaction between each pair of atom types—best achieved by minimizing the number of different types defined. In practice, the definition of atom types must reflect a compromise between these two conflicting influences.

A further consideration is that the method must be capable of automatically atom typing the wide diversity of compounds that are found as ligands in the PDB. This contrasts with the work of Melo and Feytmans,[25] who were only required to define atom types for those atoms found in the 20 common amino acid residues. Many different methods for automatic atom typing have been proposed, such as that of Hendlich et al.[30] Because the problem of sparse data restricts the number of atom types we can define, a somewhat less elaborate method is more appropriate for our purposes.

TABLE I.
The 351 PDB Entries Used to Generate BLEEP-1.

148l	152l	154l	183l	1aba	1abo	1adl	1ads	1aiz	1aky
1alk	1ami	1amj	1amt	1aoc	1aoz	1aph	1apm	1arc	1aru
1asw	1bbh	1bcx	1bel	1ben	1bfb	1bhp	1bic	1bkf	1bmd
1btl	1btn	1bvc	1byb	1cag	1can	1caz	1cb2	1cbn	1cdg
1ceg	1cel	1cfb	1chm	1cka	1cle	1cll	1cmb	1cmc	1cmp
1cnx	1coy	1csn	1csr	1ctf	1ctj	1cyd	1cyo	1daa	1dad
1dag	1dbs	1ddt	1det	1dif	1dkx	1dmb	1dor	1dpg	1drb
1drf	1dyr	1eas	1eau	1ecf	1emd	1eno	1epm	1epn	1eta
1fdx	1fel	1fil	1fiv	1fkg	1flr	1fmb	1fmc	1fnb	1fnc
1frb	1frd	1frp	1fua	1gaf	1gah	1gar	1gd1	1gdo	1ghb
1gia	1gky	1gma	1gmp	1gnr	1gof	1gpb	1gra	1grg	1gsa
1gse	1han	1hck	1hfc	1hgx	1hiv	1hml	1hne	1hnl	1hpm
1hsb	1hsl	1hug	1hur	1hxn	1hxp	1iag	1ida	1igs	1inc
1iso	1isu	1ivd	1jrs	1kap	1kel	1knt	1lam	1lcf	1lcp
1lec	1len	1les	1lic	1lin	1lkl	1llo	1lma	1lmc	1lob
1loc	1luc	1lzb	1mai	1mbd	1mdc	1mdl	1mfa	1mka	1mrk
1mtr	1mzm	1nba	1nci	1nco	1nfp	1nhk	1nhp	1nic	1nnc
1oen	1olb	1opb	1orb	1ova	1oyb	1p02	1pbe	1pbp	1pca
1pch	1pda	1phc	1phe	1php	1pii	1pip	1pk4	1pnf	1pob
1pot	1ppa	1ppf	1ppk	1ppl	1ppn	1ppp	1puc	1ras	1rbw
1rca	1rcd	1rcf	1rdn	1rie	1rn1	1rn4	1rnc	1rnn	1rpg
1rsm	1rsy	1rtm	1rtu	1rza	1rzd	1rze	1s01	1sac	1sbp
1scn	1sdk	1sgc	1sgp	1slf	1slg	1slt	1sre	1sty	1sub
1sup	1svb	1tad	1tca	1tew	1tgs	1tgt	1thb	1the	1thm
1thw	1tif	1tla	1tlm	1tml	1tnh	1top	1tpf	1tph	1tpp
1tsd	1tyr	1tys	1ubs	1udg	1udh	1urn	1vhh	1vid	1vpt
1wap	1wht	1xan	1xic	1xif	1xih	1xnb	1xzb	1xzl	1ycc
1ydb	1ymc	207l	256b	2abk	2acq	2acr	2acs	2ak3	2alp
2cbc	2ccy	2cmd	2cst	2ctc	2cut	2cwg	2cy3	2cyp	2dnj
2dri	2erl	2gmt	2gst	2hft	2hmq	2hts	2ilk	2imn	2lig
2mb5	2mcm	2mhr	2mlt	2msb	2nad	2ohx	2pgd	2pia	2por
2prd	2ran	2rox	2sn3	2tci	2tmn	2trx	2wrp	3chy	3cla
3cyh	3dfr	3ebx	3grs	3rn3	3rnt	3tmn	3tpi	4azu	4bp2
4dfr	4enl	4fgf	4lzm	4upj	5cna	5fd1	5p21	5pti	5tim
5tmn	6ebx	6est	6ldh	7cpp	7gch	7rsa	8est	8rxn	8tln
9ldt									

We have assigned the atom types by means of our SATIS scheme. This approach, which is described in more detail elsewhere,[27] is based on the list of covalently bonded partners for each atom, with no explicit ''bond type.'' The information for each atom is formulated as a ten-digit connectivity code. The first two digits are the atom's atomic number (e.g., **06** for carbon or **16** for sulfur). The remainder of the code consists of four two-digit numbers representing the atomic numbers of the atom's covalently bonded partners, in ascending numerical order. If an atom has fewer than four bonded partners, the remaining positions in the connectivity code are filled with **99**. As examples, a peptide nitrogen would have the connectivity code **0701060699** and a water oxygen atom **0801019999**. An atom type comprises one or more connectivity codes. The grouping together of several connectivity codes into single atom types, required to reduce the problem of sparse data, is arbitrary, but designed to reflect chemical similarities. The groupings we have chosen to use in generating BLEEP are shown in Table III. The choice inevitably reflects some compromise of the chemical discrimination of the atom types, in some cases averaging across a range of chemical environments to obtain reasonable quantities of data for each pair.

Given that the success of Jones et al.[31] in predicting ligand binding was based largely on correct prediction of hydrogen bonding and other polar interactions, we felt that an adequate de-

TABLE II.
The 188 PDB Entries Used to Generate BLEEP-2.

152l	154l	183l	1aba	1adl	1ads	1aky	1arc	1aru	1asw
1bcx	1bel	1bfb	1bhp	1bic	1bkf	1btl	1btn	1bvc	1can
1caz	1cbn	1ceg	1cfb	1cll	1cmp	1cnx	1csn	1ctf	1ctj
1cyo	1dad	1dag	1dbs	1det	1dif	1dmb	1drf	1dyr	1eas
1eau	1emd	1eno	1fdx	1fel	1fil	1fkg	1fmb	1fnb	1fnc
1frb	1frd	1fua	1gia	1gky	1gma	1gnr	1gsa	1han	1hck
1hfc	1hml	1hnl	1hpm	1hug	1hxn	1iag	1igs	1inc	1ivd
1knt	1lec	1lic	1lin	1llo	1lma	1lmc	1lzb	1mai	1mbd
1mdc	1mdl	1mrk	1mzm	1nfp	1nnc	1orb	1oyb	1pbe	1pbp
1pca	1pch	1pda	1phc	1phe	1php	1pk4	1pnf	1ppa	1ppn
1ppp	1puc	1ras	1rbw	1rca	1rcd	1rcf	1rie	1rn4	1rnc
1rpg	1rsm	1rsy	1rtu	1rza	1rzd	1rze	1s01	1sbp	1sgc
1sty	1sub	1sup	1tca	1tew	1tgt	1thm	1thw	1tif	1tla
1tml	1tnh	1top	1tpp	1tys	1udg	1udh	1vhh	1vid	1xnb
1xzb	1xzl	1ycc	1ydb	1ymc	2abk	2acq	2acr	2acs	2alp
2cbc	2cmd	2ctc	2cut	2cy3	2dri	2erl	2gmt	2hft	2hts
2ilk	2imn	2mb5	2mcm	2mhr	2mlt	2pia	2por	2prd	2ran
2sn3	3chy	3cla	3dfr	3ebx	3rn3	3rnt	4bp2	4fgf	4lzm
5fd1	5p21	5pti	6est	6ldh	7cpp	7gch	7rsa		

scription of hydrogen bonding would be essential to our method. To this end, we decided to include polar hydrogens (those bonded to oxygen or nitrogen) as interaction sites in BLEEP. The polar hydrogens were positioned using standard bond lengths and angles, their coordinates being calculated by the program HBPLUS.[32] The inclusion of the proton ... acceptor interaction, as well as the donor ... acceptor one, is designed to ensure that good hydrogen bonding geometries are given favorable energies by the potential. The total numbers of atom types defined for our dataset were: hydrogen, 2; nonmetals, 38; and metals, 18.

Generation of Distance Distributions

The BLEEP-1 dataset consisted of 440 protein chains, together with the ligands bound to them, whereas that for BLEEP-2 had 222 such chains. In the latter case, we wished also to include interactions with water. If one uses only the PDB data, one will include explicitly only those waters that are found in the PDB files. This will not take account of all the water present, even in the crystal phase. Although waters trapped in cavities or involved in internal hydrogen bonding may be seen if they have long residence times, waters at the interface between the bulk solvent and the protein surface are unlikely to appear, even in high-resolution X-ray structures.[33] This means that atoms that are, in reality, exposed to the solvent may appear in the data to be ''seeing'' empty space. To correct for this, we used the program AQUARIUS2[34,35] to generate the ''missing'' waters in the first hydration shell.

BLEEP-2 was based on 188 PDB entries (Table II), almost all single-chain proteins, taken from the larger dataset. The BLEEP-2 dataset was designed to avoid taking chains out of the context of their multimers, as the solvation algorithm used would have generated waters around atoms that are in fact in hydrophobic contact with another protein chain.

Our distance distributions are based on all the atom–atom distances for those pairs of atoms shown in Table IV. For BLEEP-1, that is those contacts that include an atom from the specified ligand type (ligand_1 in Table IV) in the set of 351 PDB structures. For BLEEP-2, interactions involving (crystallographic or AQUARIUS-derived) water molecules are also included, and the set of 188 structures is used. Protein–protein interactions, or those between protein and ''other'' ligands (ligand_2 in Table IV), are always excluded. Interactions involving metals are considered separately, using their own dataset[28] and one atom type per element.

Each atom–atom distance of 8.0 Å or less is recorded and included in the distribution corresponding to the two atom types involved. There is only one distribution for each pair of atom types,

TABLE III.
Grouping of Connectivity Codes into Atom Types for H, C, N, O, and S.

Type	Connectivity Codes	Description
0101	**0107999999**	H bonded to N
0102	**0108999999**	H bonded to O
—	**0106999999, 0116999999,** etc.[b]	Not considered in potential
0600	**0601050607,** etc.	C with unusual bonds/partners
0601	**0601010101, 0601010106, 0601010606, 0601060606, 0606060606**	C nonpolar, saturated
0602	**0601010699, 0601060699, 0601069999, 0606060699, 0606069999**	C nonpolar, unsaturated/aromatic
0603	**0601010107, 0601010607, 0601060607, 0606060607, 0601010115, 0601010615, 0601060615, 0606060615, 0601010133,** etc.	C bonded to N/P/As, saturated
0604	**0601010108, 0601010608, 0601060608, 0606060608, 0601010609, 0601060609, 0601010617, 0601060617, 0601010635,** etc.	C bonded to O/halogen, saturated
0605	**0601070899, 0606070899**	C amide/peptide[d]
0606	**0601080899, 0606080899, 0608080899**	C carboxylate/acid/ester
0607	**0601010799, 0601060799, 0606060799, 0606079999, 0607999999, 0601011599, 0601061599, 0606061599, 0606159999, 0615999999**	C bonded to N/P and unsaturated/aromatic
0608	**0601010899, 0601060899, 0606060899, 0606089999, 0608999999, 0606060999, 0606061799, 0606063599, 0606065399,** etc.	C bonded to O/halogen and unsaturated/aromatic
0610[a]	**0601010116, 0601010616, 0601060616, 0601060716, 0601060816, 0601061699, 0606060616, 0606061699, 0606071699,** etc.	C bonded to S/Se
0612	**0601010707, 0601060707, 0606060707, 0601070707, 0606070707, 0601010808, 0601060808, 0606060808, 0601080808,** etc.	C bonded to multiple polar atoms, saturated
0613	**0601070799, 0606070799, 0607070799, 0607079999, 0607070899, 0607080899, 0607089999,** etc.	C bonded to multiple polar atoms, unsaturated / aromatic *not* 0605, 0606
0617	**0607799999, 0601010180, 0601010182,** etc.	C bonded to metal
0701	**0701010199, 0701010699, 0706060699**	N with 3 bonds, nonpolar *not* 0702
0702	**0701060699**	N peptide or secondary amine
0703	**0706069999, 0706999999**	N acceptor, including cyanide
0704	**0701010101, 0701010106, 0701010606, 0701060606, 0706060606**	N with 4 bonds (charged)
0706	**0701011699, 0710060799, 0701060899, 0701061699, 0710070799, 0701151599, 0706060799, 0706079999, 0706080899,** etc.	N bonded to nonpolar atom(s)
0708	**0706062699, 0707269999,** etc.	N bonded to metal
0800	**0801999999, 0801059999, 0805999999, 0808999999,** etc.	O with unusual bonding/partners
0801	**0801019999**	O water
0802	**0801069999**	O hydroxyl
0803	**0806999999**	O in O=C[d]
0804	**0806069999**	O ether
0805	**0801159999, 0806159999, 0815159999, 081599999,** etc.	O bonded to P/As
0806	**0801169999, 0806169999, 0816999999,** etc.	O bonded to S
0807	**0801079999, 0807999999,** etc.	O bonded to N
0808	**0823999999, 0826269999, 0829299999, 0829999999,** etc.	O bonded to metal
1601	**1601069999, 1606060699, 1606069999, 1606169999,** etc.	S reduced
1602	**1606060808, 1606060899, 1606070808, 1606080808,** etc.	S oxidized
1603	**1626262699, 1626269999,** etc.	S bonded to metal

[a] For historical reasons, some type designations (e.g., *0609* or *0705*) are missing. The types originally corresponding to these have been merged with others.
[b] The use of ''etc.'' indicates that other connectivity codes not listed are also assigned to this type. These can be inferred from the Description column.
[c] For other elements (atomic number *xx*), there is a single atom type *xx01* comprising all codes starting with *xx*.
[d] In other work,[27] we defined extension codes for SATIS to distinguish between different C=O containing functional groups. These do not affect the definition of atom types for BLEEP, and are therefore neither described in the text nor listed in the table.

so the distribution for types *0601* (alkyl carbon) and *0702* (peptide-like nitrogen), for example, incorporates cases in which the carbon is in the ligand with the nitrogen in the protein and vice versa. The distributions are stored using bins of width 0.1 Å, considered to be an appropriate gran-

TABLE IV.
Interactions Included in the Potential.

Interactions Included	BLEEP-1	BLEEP-2
$Ligand_1$—protein	✓	✓
$Ligand_1$—water	✓	✓
$Ligand_1$—$ligand_2$	✓	✓
Protein—water	×	✓
$Ligand_2$—water	×	✓
Water—water	×	✓
$Ligand_2$—protein	×	×
Protein—protein	×	×

ularity for these data. The 40 nonmetal atom types give rise, in principle, to 820 pair distributions. In the BLEEP-1 dataset (351 structures), 220 of 820 pairs give no data at all and another 202 pairs give fewer than 50 observations. For BLEEP-2 (188 structures), there are 276 pairs with no data and another 193 with fewer than 50 distances. While this may, at first sight, appear problematic, we would argue that the pairs rarely or never observed are precisely those rarely or never required for a description of protein–ligand interactions. Because BLEEP has been generated using a large dataset of protein–ligand complexes, we expect the overwhelming majority of atom–atom pairs in such complexes to be well described by it.

The inclusion of water vastly increases the quantity of data. Despite incorporating only around half the number of proteins used for BLEEP-1, the water-inclusive BLEEP-2 is based on more than eight times as many atom–atom distances (5,583,239 from 188 structures, as opposed to 676,538 from 351 structures).

Calculation of Potentials and Reference State

We developed software to convert the distance distributions into pair potentials using the methodology of Sippl,[5] which is essentially an interpretation of Boltzmann statistics. This was performed for all 820 possible pairs of the 40 nonmetallic atom types in our library.

The mean force potential calculated for each pair is in fact the difference between the potential for that pair and a reference potential. The reference potential is that which would be obtained by using the distance distribution for the whole dataset, rather than just for one pair of types, and thus reflects the average interaction between atom-type pairs. To generate a usable potential, it is therefore necessary to calculate the reference potential, or to obtain an approximation to it.

This is, unfortunately, a nontrivial task. We have looked at two contrasting approaches to this question. The first is that suggested by Bahar and Jernigan.[14] Their procedure involves a reverse-Boltzmann fitting of the overall data, similar to that used for the individual distributions. However, two new problems arise in connection with this absolute potential, as opposed to the previous relative ones. The less serious one is calibration; this is achieved by setting the energy at the longest distance considered (8.0 Å) to zero. A more serious issue is normalization. For the individual pair difference potentials, normalization on the basis of available volume is unnecessary, because the appropriate terms from the overall and particular distributions cancel out. For the overall potential, however, we must consider the available volume at any given distance, r, from the center of an atom. For interacting spherical atoms *in vacuo*, this is given simply by $4\pi r^2 \delta r$. Applying this here, however, gives an average potential that is weakly repulsive over the entire distance range with no attractive region at typical interaction distances. This is probably due to the available volume for atoms in protein–ligand complexes deviating significantly from r^2 proportionality, especially at short range, due to excluded volume effects.

The second approach to the question of the reference potential is a more pragmatic one. We know that the short-range repulsive region of the potential is not sampled by the data. Given our expectation that the average energy will have a repulsive wall close to the sum of typical van der Waals radii, we simply import such a function from elsewhere and use it as the reference potential. Because BLEEP largely describes first-row atoms, the neon–neon interaction potential of Ng et al.,[36] shown in Figure 1, seems a reasonable choice for this purpose. Its short-range repulsive wall takes account of a region of the potential not sampled by the interaction data, whereas its attractive well is extremely shallow (minimum −0.36 kJ/mol at 3.09 Å). Its absolute values at typical interaction distances, and also at longer range, are generally small compared with those of the PMF pair potentials. This approach might, however, be criticized for ''mixing and matching'' mean force statistical potentials with theoretical semiempirical ones. A related comment is that one would be mixing two quantities which, although both ''en-

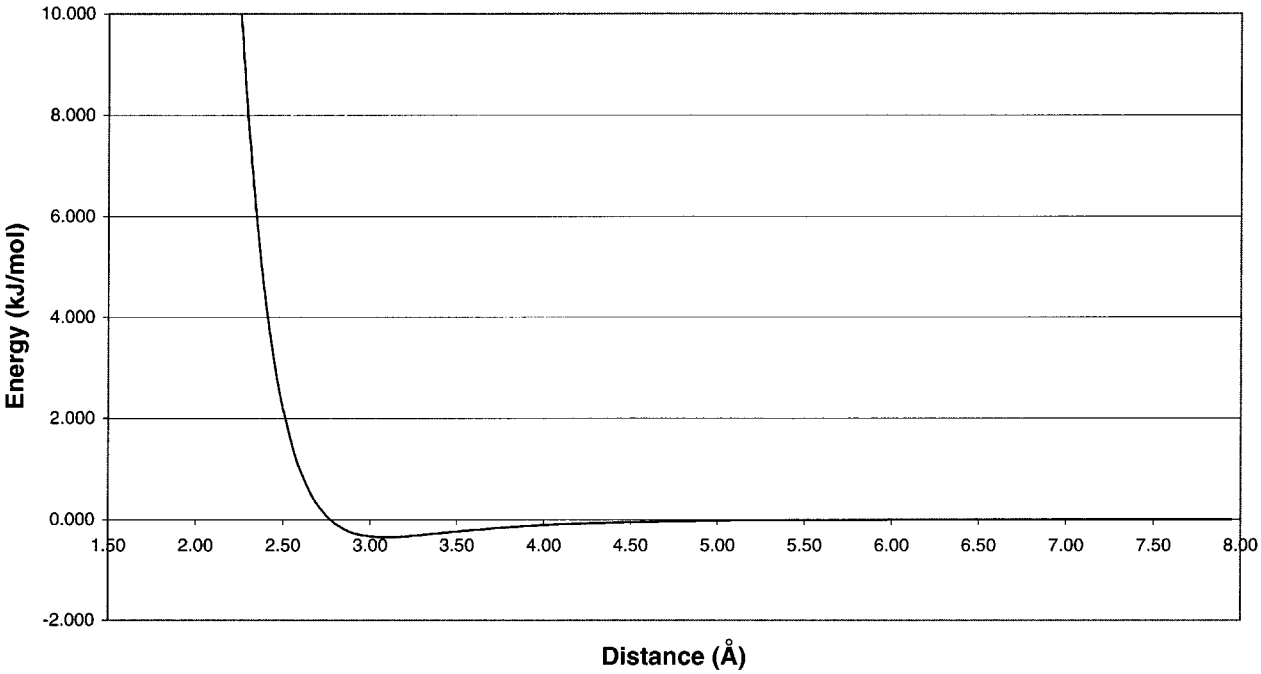


FIGURE 1. Imported average function for BLEEP-2. This is the neon–neon potential of Ng et al.[36] Its repulsive wall around the sum of typical van der Waals radii takes account of a region of the potential not sampled by the interaction data. The shallowness of its well (minimum −0.36 kJ/mol at 3.09 Å) ensures that the shapes of the pair potentials are dominated by energetic effects from the PMF term.

ergies,'' are not the same thing in thermodynamic terms. The alternative view is that the imported reference potential should be thought of simply as a set of adjustable parameters in what is clearly an empirical method. We have taken this approach with BLEEP. For pairs involving hydrogen, we displace this potential by 1 Å so that the repulsive wall occurs at shorter distances, as discussed in what follows. For atom-type pairs that include hydrogen, this function is displaced to distances 1.0 Å shorter; that is:

$$\phi^{\mathrm{H}}(r) = \phi(r + 1.0\ \text{Å}) \qquad (1)$$

where $\phi^{\mathrm{H}}(r)$ and $\phi(r)$ are the imported reference potentials at distance r for pairs respectively including and excluding hydrogen.

Sippl's method of calculating potentials from distance distributions uses a weighting function (σ) to combine the sample data for a given pair with the signal from the average interaction. The potential of mean force[5] for atom types a and b at distance r is given by:

$$\Delta \mathrm{E}^{ab}(r) = \mathrm{kT}\ln[1 + \mathrm{m}^{ab}\sigma] - \mathrm{kT}\ln[1 + \mathrm{m}^{ab}\sigma\{\mathrm{g}^{ab}(r)/\mathrm{f}(r)\}] \qquad (2)$$

where m^{ab} is the total number of contacts between atom types a and b, $\mathrm{g}^{ab}(r)$ is the proportion of these contacts occurring at distance r, $\mathrm{f}(r)$ is the proportion of all contacts for all types occurring at distance r, k is the Boltzmann constant, and T is the absolute temperature. As we use a histogram-based representation, r here effectively refers to a given bin of width 0.1 Å. The potentials are smoothed by the use of a moving window encompassing five adjacent bins (see later).

We set σ to 0.02, so that 50 observations are required for the observed data to carry an equal weight to the overall average of all pair distributions. A pair with no data at all would therefore be represented by the average potential for all atom–atom interactions in the dataset. On the other hand, a pair with a large quantity of data (such as *0601* and *0601* with 11,116 occurrences for BLEEP-1 and 4513 for BLEEP-2) would be represented almost entirely by its own data.

The 820 pair potentials were calculated with a granularity of 0.1 Å, using a moving window of width 0.5 Å (with weights of 1:2:4:2:1 for bins $[i-2]$ to $[i+2]$). The distance range was 1.5 to 8.0 Å (but distances below 2.5 Å were rejected as unrealistically short for pairs not involving hydro-

gen; the repulsive potential in this region comes from the imported reference function). The energy was calculated for each increment of 0.1 Å using eq. (2) and the moving window. This was carried out independently for BLEEP-1 and BLEEP-2.

Results and Discussion

Four of the 820 pairs are chosen as representative examples for the potentials. For reasons explained in what follows, we concentrate on BLEEP-2. Examples of the calculated potentials (for BLEEP-2), including the imported reference function, are shown in Figure 2. The potential for types *0702* (peptide nitrogen) and *0803* (carbonyl oxygen) (Fig. 2a), shows a minimum just below 3.0 Å, corresponding to a typical hydrogen bonding distance between these atom types. The potential, which is based on 1988 observed interactions, then becomes repulsive at somewhat longer distances. The potential for atom types *0101* (hydrogen bonded to nitrogen) and *0801* (water oxygen) is shown in Figure 2b. This shows a very strong minimum corresponding to hydrogen bonding around 2.0 Å. The potential is weakly repulsive around 4.0 Å, but slightly attractive around 5.5 Å. In this case, the number of observations (771,031) is sufficiently large, so that we can be confident that the fine details of the potential are meaningful. Figure 2c shows that the same hydrogen atom type (*0101*) has a repulsive potential with saturated alkyl carbon type *0601* at all distances up to 4.0 Å, as would be expected for atom types that are not likely often to be in close contact (based on 5933 observed distances). Similarly, we see from Figure 2d that type *0601* carbons have a repulsive potential with themselves at short distances (up to 4.0 Å), but a slightly attractive one between 4.0 and 7.0 Å (4513 observations).

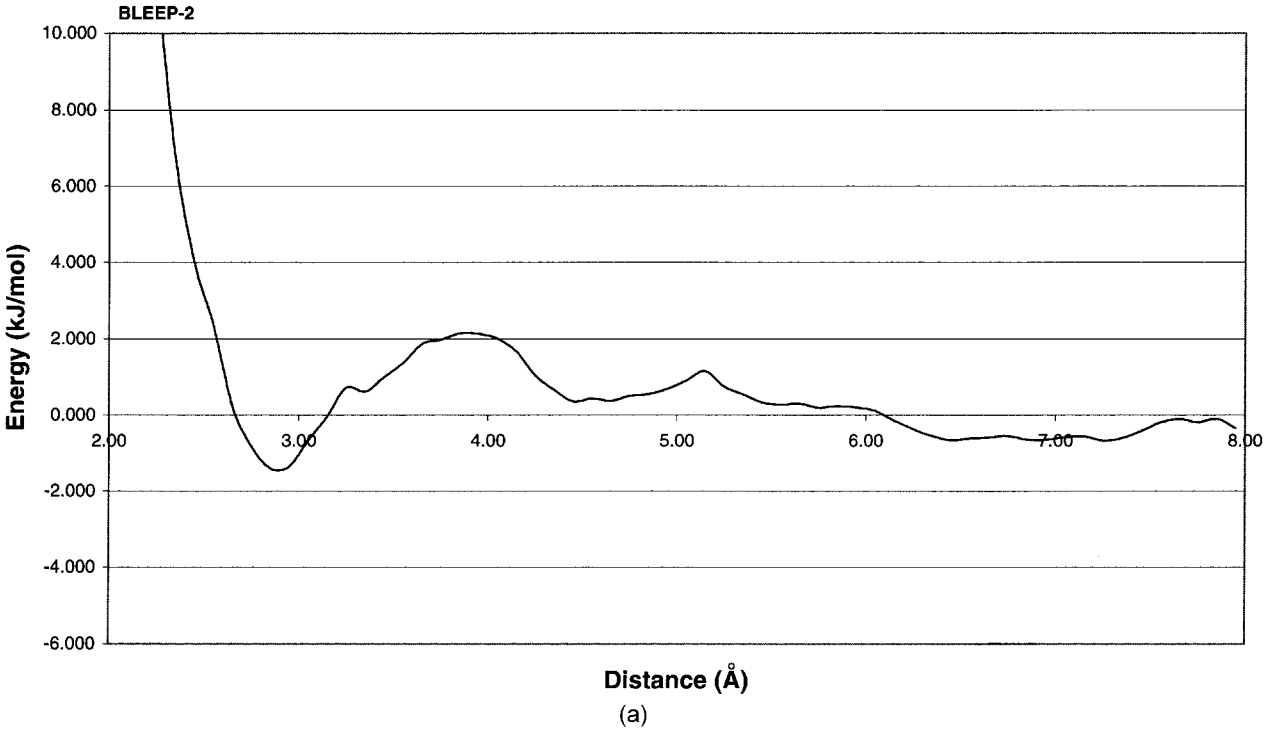
The shapes of these pair potentials are very encouraging. In general, we see short-range minima where expected for polar and hydrogen bonding interactions. Pairs of nonpolar atoms tend to prefer somewhat longer distances. The polar–hydrophobic pair (*0101–0601*) is notably repulsive at short range and is nowhere significantly attractive. These forms are in accordance with our expectations from chemical intuition and, along with the reasonably smooth form of the functions, give good grounds for expecting that real protein–ligand interactions will be successfully described by them. All these potentials are close to zero as they approach the cutoff value of 8.0 Å. The (*0702–0803*) pair potential is particularly interesting, as it can be compared with the protein backbone N...O potential calculated by Sippl[6] (which incorporates only pairs separated in sequence by at least nine residues). The form of our potential is generally very similar to that of Sippl's. Both potentials show a minimum just below 3.0 Å, with an energy barrier separating this from the near-zero long-range interaction. The peaks close to 4.0 Å in each case are entropic in origin. Because our data correspond to protein–ligand pairs, rather than to atoms connected by a protein chain, it is not surprising that there are some differences in the detail of the peak shapes. Another contribution to these differences probably comes from the rather broad range of atoms incorporated in our *0803* oxygen atom type (see Table III). The (*0101–0801*) pair potential has a similar overall shape, as expected for frequently hydrogen-bonded atoms, but a much deeper minimum in the attractive region and a smaller barrier separating this from the long-range part of the potential.

A number of preliminary test calculations were performed using BLEEP-1 and BLEEP-2. The most important of these involved calculating the interaction energies of the nine serine proteinase–inhibitor complexes given by Zhang et al.[37] The correlation coefficients between calculated and experimental binding energies were 0.60 for BLEEP-1 and 0.71 for BLEEP-2. This result, along with a number of others (data not shown), suggested to us that BLEEP-2 was the more promising of our two protocols and led us to concentrate mainly on it. A detailed account of the calculation and validation of interaction energies using BLEEP is given in the accompanying paper.[24]

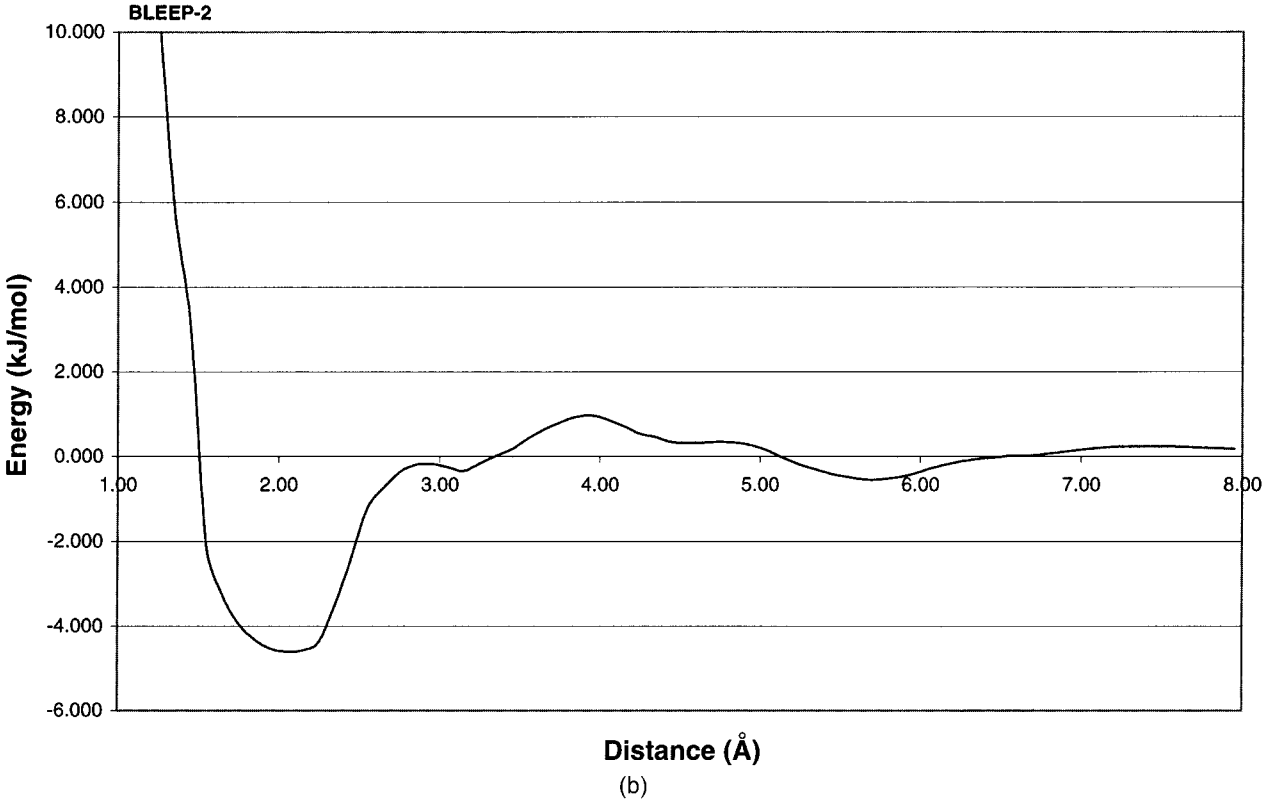
Conclusions

We have successfully generated BLEEP, a PMF describing protein–ligand interactions, using a broad-based sample of high-resolution X-ray structural data from the Brookhaven PDB. BLEEP is generally applicable to ligand molecules of diverse size and chemical composition, and represents a significant advance on previous methods[18] that were focused on a single class of molecules.

Taking account of homology, we were able to produce a representative dataset and from it to generate distance distributions of protein–ligand pair interactions for all 820 pairs of 40 carefully

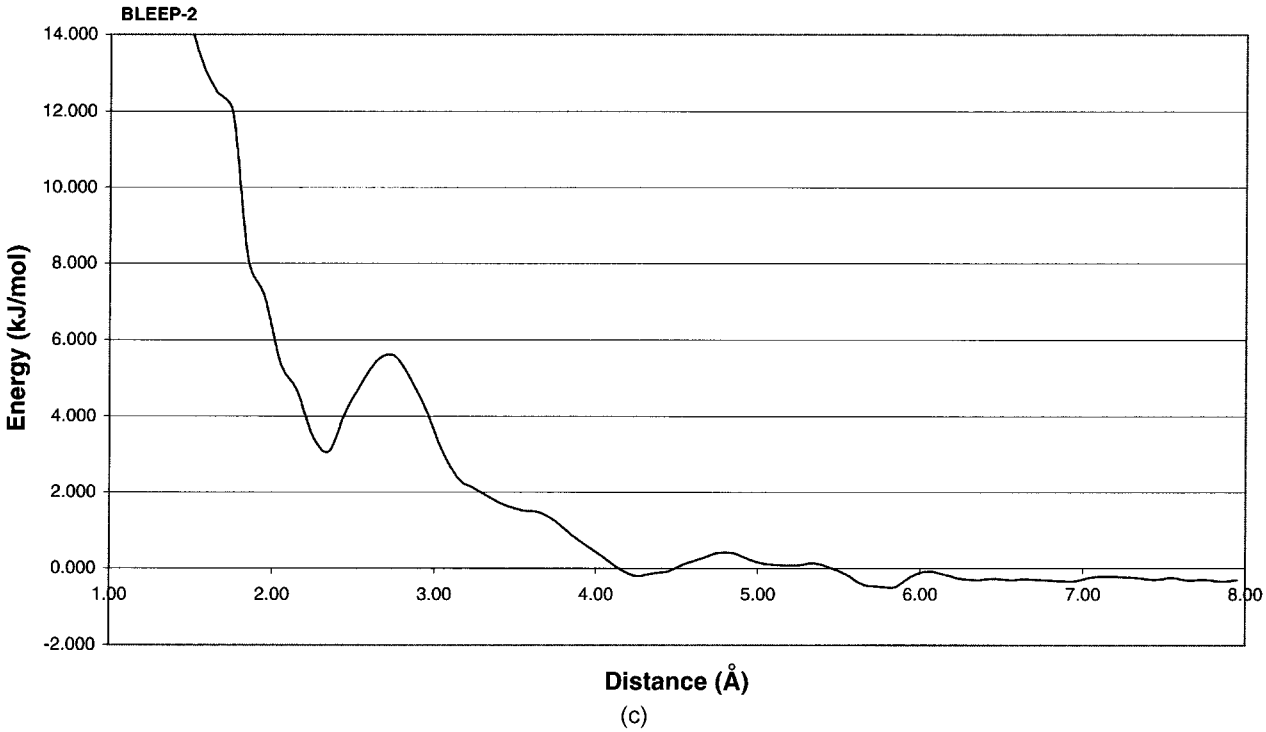


(a)

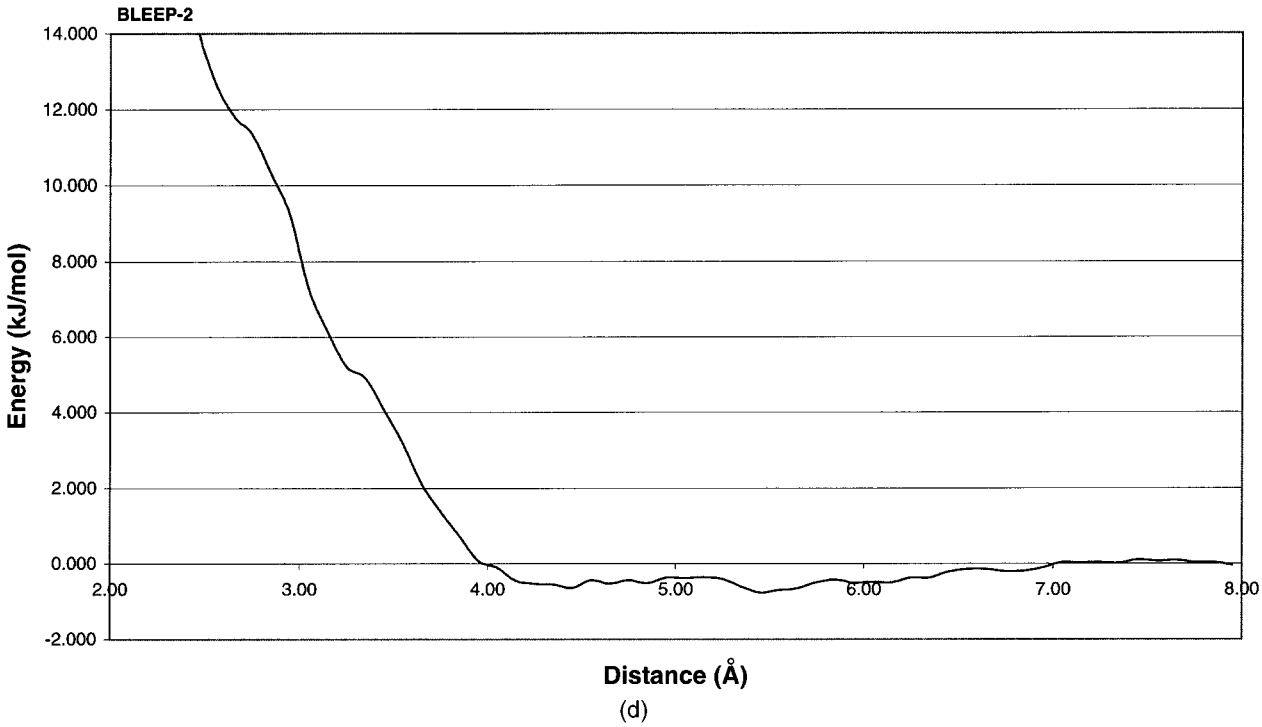


(b)

FIGURE 2. (a) BLEEP-2: atom types *0702* (N) and *0803* (O). (b) BLEEP-2: atom types *0101* (H) and *0801* (C). (c) BLEEP-2: atom types *0101* (H) and *0601* (C). (d) BLEEP-2: atom types *0601* and *0601.*



(c)



(d)

FIGURE 2. Continued.

chosen atom types. Assuming Boltzmann energetics, we were able to convert these distributions into energy-like functions between 2.5 Å and 8.0 Å (or 1.5 Å and 8.0 Å where hydrogen was involved). An assumed average energy function was imported to provide a realistic repulsive potential at very short range. Two versions of the dataset and pair potentials were produced, one excluding

(BLEEP-1) and one including (BLEEP-2) interactions with water. The form of the potentials was in accord with chemical intuition and they were close to zero near the cutoff distance of 8.0 Å. Preliminary calculations strongly suggest that the water-inclusive version, BLEEP-2, is the more effective.

We believe that BLEEP is suitable for the description of general protein–ligand interactions, and particularly so in the context of computer-aided drug design. The use of BLEEP to calculate protein–ligand interaction energies is described in detail in part II of this study.[24]

Acknowledgments

This is a publication from the BBSRC Structural Biology Centre in Birkbeck College and University College London.

References

1. 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. J Mol Biol 1977, 112, 535.
2. Sippl, M. J. Curr Opin Struct Biol 1995, 5, 229.
3. Mitchell, J. B. O.; Price, S. L. J Comput Chem 1990, 11, 1217.
4. Marelius, J.; Hansson, T.; Åqvist, J. Int J Quantum Chem 1998, 69, 77.
5. Sippl, M. J. J Mol Biol 1990, 213, 859.
6. Sippl, M. J. J Mol Biol 1996, 260, 644.
7. Sippl, M. J.; Ortner, M.; Jaritz, M.; Lackner, P.; Flöckner, H. Fold Design 1996, 1, 289.
8. Jones, D. T.; Taylor, W. R.; Thornton, J. M. Nature 1992, 358, 86.
9. Reva, B. A.; Finkelstein, A. V.; Sanner, M. F.; Olson, A. J. Prot Eng 1997, 10, 865.
10. O'Donoghue, S. I.; Nilges, M. Fold Des 1997, 2, S47.
11. Vajda, S.; Sippl, M.; Novotny, J. Curr Opin Struct Biol 1997, 7, 222.
12. Sippl, M. J. Proteins 1993, 17, 355.
13. Melo, F.; Feytmans, E. J Mol Biol 1998, 277, 1141.
14. Bahar, I.; Jernigan, R. L. J Mol Biol 1997, 266, 195.
15. Thomas, P. D.; Dill, K. A. J Mol Biol 1996, 257, 457.
16. Sunyaev, S. R.; Eisenhaber, F.; Argos, P.; Kuznetsov, E. N.; Tumanyan, V. G. Proteins 1998, 31, 225.
17. Zhang, L.; Skolnick, J. Prot Sci 1998, 7, 112.
18. Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Prot Eng 1995, 8, 677.
19. Verkhivker, G. M.; Rejto, P. A. Proc Natl Acad Sci USA 1996, 93, 60.
20. Rejto, P. A.; Verkhivker, G. M. Proteins 1997, 28, 313.
21. Wallqvist, A.; Jernigan, R. L.; Covell, D. G. Prot Sci 1995, 4, 1881.
22. DeWitte, R. S.; Shakhnovich, E. I. J Am Chem Soc 1996, 118, 11733.
23. DeWitte, R. S.; Ishchenko, A. V.; Shakhnovich, E. I. J Am Chem Soc 1997, 119, 4608.
24. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M.; Thornton, J. M. J Comput Chem (this issue).
25. Melo, F.; Feytmans, E. J Mol Biol 1997, 267, 207.
26. Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. Structure 1997, 5, 1093.
27. Mitchell, J. B. O.; Alex, A.; Snarey, M. J Chem Inf Comput Sci (accepted for publication).
28. Further details of the datasets used in this work are available on the internet at http://www.biochem.ucl.ac.uk/bsm/biocomp/mitchell_etal_list.html.
29. Laskowski, R. A.; Hutchinson, E. G.; Michie, A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. Trends Biochem Sci 1997, 22, 488.
30. Hendlich, M.; Rippmann, F.; Barnickel, G. J Chem Inf Comput Sci 1997, 37, 774.
31. Jones, G.; Willet, P.; Glen, R. C.; Leach, R.; Taylor, R. J Mol Biol 1997, 267, 727.
32. McDonald, I. K.; Thornton, J. M. J Mol Biol 1994, 238, 777.
33. Karplus, P. A.; Faerman, C. Curr Opin Struct Biol 1994, 4, 770.
34. Pitt, W. R.; Murray-Rust, J.; Goodfellow, J. M. J Comput Chem 1993, 14, 1007.
35. Goodfellow, J. M.; Pitt, W. R.; Smart, O. S.; Williams, M. A. Comput Phys Commun 1995, 91, 321.
36. Ng, K.-C.; Meath, W. J.; Allnatt, A. R. Molec Phys 1979, 37, 237.
37. Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. J Mol Biol 1997, 267, 707.