PROTEINS: Structure, Function, and Genetics 43:89–102 (2001)

Residue Frequencies and Pairing Preferences at Protein–Protein Interfaces

Fabian Glaser,[1] David M. Steinberg,[2] Ilya A. Vakser,[3] and Nir Ben-Tal[1]*

[1]*Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel*
[2]*Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv, Israel*
[3]*Department of Cell and Molecular Pharmacology, Medical University of South Carolina, Charleston, South Carolina*

ABSTRACT **We used a nonredundant set of 621 protein–protein interfaces of known high-resolution structure to derive residue composition and residue–residue contact preferences. The residue composition at the interfaces, in entire proteins and in whole genomes correlates well, indicating the statistical strength of the data set. Differences between amino acid distributions were observed for interfaces with buried surface area of less than 1,000 Å^2 versus interfaces with area of more than 5,000 Å^2. Hydrophobic residues were abundant in large interfaces while polar residues were more abundant in small interfaces. The largest residue–residue preferences at the interface were recorded for interactions between pairs of large hydrophobic residues, such as Trp and Leu, and the smallest preferences for pairs of small residues, such as Gly and Ala. On average, contacts between pairs of hydrophobic and polar residues were unfavorable, and the charged residues tended to pair subject to charge complementarity, in agreement with previous reports. A bootstrap procedure, lacking from previous studies, was used for error estimation. It showed that the statistical errors in the set of pairing preferences are generally small; the average standard error is ≈0.2, i.e., about 8% of the average value of the pairwise index (2.9). However, for a few pairs (e.g., Ser–Ser and Glu–Asp) the standard error is larger in magnitude than the pairing index, which makes it impossible to tell whether contact formation is favorable or unfavorable. The results are interpreted using physicochemical factors and their implications for the energetics of complex formation and for protein docking are discussed. Proteins 2001;43:89–102.** © **2001 Wiley-Liss, Inc.**

Key words: molecular recognition; protein modeling; protein docking; knowledge-based potentials; statistical energy functions

INTRODUCTION

The analysis of protein–protein interfaces in search of indicators of binding sites in proteins began during the 1970s.[1,2] The interfaces were found to have more hydrophobic residues than the rest of the protein surface. Recent large-scale studies of protein complexes[3–9] have confirmed the importance of hydrophobicity in protein–protein interactions and tested the importance of other factors, such as interface propensity of the residues, accessible surface area, and degree of planarity and protrusion. However, none of these factors was found to dominate the data set of protein–protein complexes.[10] A subsequent test using the average of these factors as an indicator of protein binding sites showed only a 66% success rate for 59 predictions.[11] Other aspects of protein–protein complexes that have been investigated include packing and buried surface area.[3,12–15] A generally accepted conclusion is that the interacting proteins have a high degree of surface complementarity,[16] but electrostatic complementarity is observed as well.[17–19]

Protein–protein interfaces are complex environments, and deciphering the details of the interactions between residues based on physical chemistry analysis may be difficult. The alternative is to use knowledge-based approaches, i.e., to derive interaction potentials from the propensities of residues to interact with each other, as observed in the collection of known structures of protein complexes. Residue–residue contact preferences have been extensively used in studies of protein structures.[20–22] The coarse-grained models, in which residues are represented by one or two points, combined with knowledge-based residue–residue potentials, allow one to reduce the number of conformational degrees of freedom drastically and to incorporate the experimental data on the interaction of molecular fragments. Residue-based structure prediction techniques have become a practical way for building a low-resolution structure of a protein that can potentially be further refined by atom-based methods. Although most residue-based studies have been concerned with intramolecular structures, the same principles of coarse grained, low-resolution modeling apply to the prediction of intermolecular interactions.

The structure of a protein complex is generally more difficult to obtain than the structure of individual proteins. However, a statistically significant number of co-crystal-

Grant sponsor: National Science Foundation; Grant number: DBI-9808093; Grant sponsor: South Carolina NSF EPSCoR Cooperative Agreement; Grant sponsor: Israeli Ministry of Sports, Culture and Science; Grant number: 422-241.

*Correspondence to: Nir Ben-Tal, Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. E-mail bental@ashtoret.tau.ac.il

Received 21 August 2000; Accepted 30 November 2000

TABLE I. Names of the PDB Files and Subunits That Form the Interfaces Used in This Study[†]

104l-AB,	1aap-AB,	1aar-AB,	1aaz-AB,	1abr-AB,	1acb-EI,	1ade-AB,	1adu-AB,	1aer-AB,	1ahs-BC,	1aiz-AB,	1aks-AB,	1alk-AB,
1all-AB,	1anw-AB,	1aor-AB,	1aoz-AB,	1apx-AB,	1apy-AC,	1apy-AB,	1apy-BD,	1asy-AB,	1atn-AD,	1avd-AB,	1bar-AB,	1bbb-AB,
1bbh-AB,	1bbp-BD,	1bbr-EK,	1bbt-14,	1bbt-24,	1bbt-23,	1bcf-AB,	1bcm-AB,	1bdm-AB,	1bgl-AD,	1bgs-FG,	1bhm-AB,	1bin-AB,
1bjm-AB,	1blb-AB,	1bmf-AG,	1bmf-CD,	1bmf-DF,	1bmf-DG,	1bmt-AB,	1bmv-12,	1bnc-AB,	1bnd-AB,	1bov-BC,	1bpl-AB,	1bql-LY,
1bro-AB,	1brs-CF,	1bsr-AB,	1bun-AB,	1bvp-23,	1c2r-AB,	1cax-AC,	1cax-CF,	1cbi-AB,	1cdc-AB,	1cdk-AB,	1cdl-AC,	1cea-AB,
1cgj-EI,	1cgl-AB,	1chk-AB,	1chm-AB,	1cho-EI,	1chr-AB,	1cki-AB,	1cle-AB,	1clx-AB,	1cmc-AB,	1cns-AB,	1col-AB,	1cpc-AB,
1cpc-AK,	1csg-AB,	1csk-AD,	1csm-AB,	1cud-AB,	1cwe-AC,	1cwp-AB,	1cyd-CD,	1d66-AB,	1daa-AB,	1dbq-AB,	1dcp-GH,	1dea-AB,
1dek-AB,	1der-FG,	1dfj-EI,	1dfn-AB,	1dif-AB,	1dir-AB,	1dkt-AB,	1dky-AB,	1dlh-BE,	1dmx-AB,	1dnp-AB,	1dok-AB,	1dpg-AB,
1dpp-AC,	1dpr-AB,	1dsb-AB,	1dth-AB,	1dut-AB,	1dvf-BD,	1dvr-AB,	1dyn-AB,	1ebd-BC,	1ece-AB,	1ecf-AB,	1ecm-AB,	1ecp-BD,
1ecz-AB,	1edh-AB,	1edm-BC,	1efn-AB,	1efn-BD,	1efu-AC,	1efu-AB,	1efu-BD,	1epa-AB,	1ept-AB,	1ept-AC,	1ept-BC,	1esf-AB,
1etp-AB,	1ext-AB,	1fag-AC,	1fat-AB,	1fba-BC,	1fbi-QY,	1fc1-AB,	1fc2-CD,	1fcb-AB,	1fcc-AC,	1fcd-AB,	1fcd-BD,	1fia-AB,
1fie-AB,	1fin-AB,	1fjl-AB,	1fjm-AB,	1fki-AB,	1fle-EI,	1fod-12,	1fod-13,	1fos-GH,	1frp-AB,	1frt-AC,	1frt-BC,	1frv-AC,
1frv-CD,	1fss-AB,	1fug-AB,	1fuj-AB,	1fuq-AB,	1fvp-AB,	1fxi-AD,	1fxr-AB,	1gad-OP,	1gam-AB,	1gar-AB,	1gdh-AB,	1gdt-AB,
1ges-AB,	1gff-12,	1gfl-AB,	1ggg-AB,	1ghs-AB,	1gif-BC,	1gla-FG,	1glq-AB,	1glu-AB,	1got-AB,	1got-BG,	1gp1-AB,	1gpm-BD,
1gri-AB,	1gse-AB,	1gto-BC,	1gtp-BI,	1gtq-AB,	1gua-AB,	1gyl-AB,	1hav-AB,	1hbh-CD,	1hcg-AB,	1hcn-AB,	1hde-AB,	1hge-AC,
1hge-CD,	1hge-DF,	1hiw-AR,	1hjr-BD,	1hle-AB,	1hlp-AB,	1hmp-AB,	1hng-AB,	1hpc-AB,	1hpl-AB,	1hrd-BC,	1hrh-AB,	1hro-AB,
1hsa-AD,	1hsb-AB,	1hsl-AB,	1hst-AB,	1htm-DF,	1htr-BP,	1htt-AB,	1huc-AB,	1hul-AB,	1hur-AB,	1hxp-AB,	1hyh-AB,	1hyl-AB,
1ice-AB,	1ids-AC,	1ies-BE,	1igc-AH,	1igc-AL,	1ihf-AB,	1ilr-12,	1inh-AB,	1isu-AB,	1ith-AB,	1jst-AC,	1kba-AB,	1kif-BF,
1kir-BC,	1kny-AB,	1kob-AB,	1kpb-AB,	1kpt-AB,	1lcp-AB,	1leh-AB,	1lgb-AC,	1lmb-34,	1lmk-EG,	1lmw-BD,	1lpb-AB,	1lta-CE,
1lti-AG,	1lts-AC,	1lts-DE,	1luc-AB,	1lwi-AB,	1lya-AB,	1lya-BD,	1lyl-AC,	1lyn-AB,	1mac-AB,	1mas-AB,	1mco-HL,	1mda-HL,
1mda-BJ,	1mda-HJ,	1mdp-12,	1mdt-AB,	1mdy-AB,	1mec-14,	1mee-AI,	1mhl-AB,	1mhl-AC,	1mhl-CD,	1mka-AB,	1mld-AB,	1mmo-BC,
1mmo-CH,	1mmo-CE,	1mmo-DE,	1mmo-EH,	1mol-AB,	1mpm-BC,	1msa-AD,	1msp-AB,	1mtn-BF,	1mtn-FH,	1mtn-GH,	1myk-AB,	1nal-23,
1nba-AB,	1ncc-LN,	1ncc-HN,	1nci-AB,	1nco-AB,	1nfk-AB,	1nhk-LR,	1nip-AB,	1nmb-LN,	1nmb-HN,	1noy-AB,	1npo-AC,	1nsc-AB,
1nsn-HS,	1nsn-LS,	1oac-AB,	1obp-AB,	1occ-NO,	1occ-NP,	1occ-NQ,	1occ-NS,	1occ-GN,	1occ-NU,	1occ-NV,	1occ-NW,	1occ-NX,
1occ-NY,	1occ-NZ,	1occ-OQ,	1occ-OR,	1occ-OU,	1occ-OV,	1occ-OX,	1occ-PS,	1occ-PT,	1occ-PU,	1occ-PW,	1occ-QR,	1occ-QS,
1occ-QV,	1occ-QX,	1occ-QZ,	1occ-RS,	1occ-RV,	1occ-FS,	1occ-ST,	1occ-SW,	1occ-TU,	1occ-HU,	1occ-WY,	1occ-YZ,	1one-AB,
1onr-AB,	1ord-AB,	1oro-AB,	1ort-BF,	1osj-AB,	1otf-BE,	1otg-BC,	1ova-AB,	1ova-CD,	1ovo-AB,	1pag-AB,	1pam-AB,	1pbw-AB,
1pdg-AB,	1pfx-CL,	1pge-AB,	1pio-AB,	1pky-AC,	1pma-12,	1pma-CD,	1pma-CP,	1pml-BC,	1pnk-AB,	1pov-03,	1pox-AB,	1poy-12,
1pp2-LR,	1ppf-EI,	1prc-CL,	1prc-CM,	1prc-CH,	1prc-HL,	1prc-LM,	1prc-HM,	1pre-AB,	1prt-AB,	1prt-AE,	1prt-AF,	1prt-EF,
1prt-HJ,	1prt-HL,	1prt-JK,	1psa-AB,	1psd-AB,	1pvc-12,	1pvc-13,	1pvc-24,	1pvc-34,	1pvd-AB,	1pvu-AB,	1pxt-AB,	1pya-CE,
1pya-CD,	1pya-DF,	1pyi-AB,	1pyt-BD,	1pyt-AB,	1pyt-AC,	1qap-AB,	1qas-AB,	1qbe-BC,	1qor-AB,	1qpa-AB,	1qrd-AB,	1rah-BD,
1rba-AB,	1rcm-AB,	1rco-RV,	1rcp-AB,	1rdl-12,	1reg-XY,	1reg-CD,	1rfb-AB,	1rgf-AB,	1rhg-AC,	1rlb-AF,	1rn1-AC,	1rth-AB,
1rtm-12,	1rtp-23,	1rva-AB,	1rvv-12,	1sac-CD,	1sce-BD,	1sch-AB,	1scm-AB,	1scm-BC,	1scu-DE,	1scu-BE,	1seb-AB,	1seb-EH,
1sei-AB,	1sem-AB,	1set-AB,	1sft-AB,	1sgp-EI,	1slt-AB,	1slu-AB,	1smn-AB,	1smp-AI,	1spb-PS,	1sph-AB,	1sri-AB,	1seb-EH,
1stf-EI,	1stm-BC,	1sva-23,	1tab-EI,	1taf-AB,	1tah-AC,	1tbr-KS,	1tcb-AB,	1tco-AC,	1tco-AB,	1tco-BC,	1tcr-AB,	1tgx-AB,
1the-AB,	1thj-BC,	1tht-AB,	1tii-AH,	1tii-AC,	1tii-EF,	1tii-CE,	1tlf-AB,	1tmc-AB,	1tme-12,	1tme-13,	1tme-23,	1tmf-13,
1tmf-14,	1tmf-23,	1tmf-24,	1tmf-34,	1tnd-AC,	1tnf-AB,	1tnr-AR,	1tph-12,	1trk-AB,	1tro-AC,	1tsd-AB,	1tsr-AB,	1tta-AB,
1tvx-BD,	1ubs-AB,	1ucy-EH,	1ucy-HK,	1udi-EI,	1umu-AB,	1una-AB,	1urn-AB,	1vcp-BC,	1vfb-AB,	1vfb-AC,	1vhi-AB,	1vmo-AB,
1vok-AB,	1vol-AB,	1vrt-AB,	1vsc-AB,	1vsg-AB,	1wap-BC,	1wdc-AC,	1wfb-AB,	1wgt-AB,	1wht-AB,	1wtl-AB,	1xik-AB,	1xim-AC,
1xso-AB,	1xva-AB,	1xxa-DF,	1xyp-AB,	1ycb-AB,	1ygp-AB,	1yha-AB,	1ypp-AB,	1ypt-AB,	1yrn-AB,	1ytf-AD,	1ytf-BD,	1ytt-AB,
1zop-AB,	256b-AB,	2abx-AB,	2ach-AB,	2adm-AB,	2afn-BC,	2bbk-HJ,	2bbk-HL,	2bbv-BC,	2bpa-13,	2btf-AP,	2ccy-AB,	2cht-DE,
2cst-AB,	2dhf-AB,	2dld-AB,	2drp-AD,	2eip-AB,	2gls-BH,	2hhm-AB,	2hip-AB,	2hmq-CD,	2hnt-CF,	2hpp-HP,	2kai-AI,	2kai-BI,
2kau-AB,	2kau-AC,	2kau-BC,	2lig-AB,	2ltn-AC,	2ltn-AB,	2mev-12,	2mev-23,	2mev-24,	2mev-34,	2mta-AC,	2mta-AH,	2mta-AL,
2nac-AB,	2pcc-AB,	2pcd-BC,	2pcd-BN,	2pcd-MP,	2pel-BC,	2phl-BC,	2pka-AB,	2pka-BY,	2plv-14,	2pol-AB,	2psp-AB,	2ptc-EI,
2rbi-AB,	2rmc-EG,	2rsl-AB,	2rsp-AB,	2scp-AB,	2spc-AB,	2tbv-AB,	2tmd-AB,	2trx-AB,	2utg-AB,	2zta-AB,	3bto-BC,	3cro-LR,
3gap-AB,	3hhr-AB,	3hhr-BC,	3ink-CD,	3ins-BD,	3lad-AB,	3mde-AB,	3mon-CE,	3mon-CD,	3mon-BD,	3pga-24,	3pmg-AB,	3sdh-AB,
3sic-EI,	4aah-AC,	4aah-CD,	4ake-AB,	4cha-AB,	4cts-AB,	4dfr-AB,	4htc-HI,	4kbp-BC,	4sbv-AB,	4sgb-EI,	4tsl-AB,	5cna-AB,
6chy-AB,	6fab-HL,	6gsv-AB,	6pfk-CD,	6q21-AB,	8atc-AB,	8cat-AB,	8ruc-EK,	8ruc-KL,	9ldt-AB			

[†]The four leftmost characters are Protein Data Bank (PDB) identifiers and the two rightmost characters are the chain designators.

lized binary protein complexes exists (several thousand, including dimers, with the number depending on the criteria and varying in different studies), permitting systematic studies of protein–protein interactions. A number of databases of protein complexes have recently been compiled and used to investigate physicochemical and structural preferences at protein–protein interfaces.[3,7,23–29] Several studies also describe databases built to study the structural factors of protein interfaces, which are artifacts of crystallization.[13,24,30]

Jernigan and coworkers[25] have shown that a wise choice of the reference state enables the description of both intra- and intermolecular interactions using the same set of potentials. Sternberg and coworkers[27] have used database-derived pair potentials in screening predicted docked complexes, and their results are encouraging; the correct dock0ing was placed within the top 12% of the ranked structures in a set of 10 complexes.

In this study, we determined the residue–residue preferences in intermolecular interactions based on a nonredundant database of 621 co-crystallized protein–protein interfaces. The database is larger than databases used in previous studies by a factor of ≥3. Unlike previous studies, we used a bootstrapping procedure for error estimation, to evaluate the statistical significance of the results. Finally, we analyzed pairs of residues and give a physicochemical interpretation of the results. We believe that our results will prove useful in distinguishing between true and false docking configurations.

METHODS

Database of Protein–Protein Complexes

A comprehensive set of binary protein–protein complexes with known three-dimensional (3D) structures was built for validation purposes (Table I) (I.A. Vakser and A. Sali, unpublished observations; http://guitar.rockefeller.edu). The set, which was derived from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb), includes predominantly oligomeric proteins, but also enzyme-inhibitor complexes, membrane proteins, and so forth. All the entries were separated into families of those with more than 30% sequence identity. A pair of chains was considered to belong to the complex if they had the same PDB four-character structure identifier and a different chain identi-

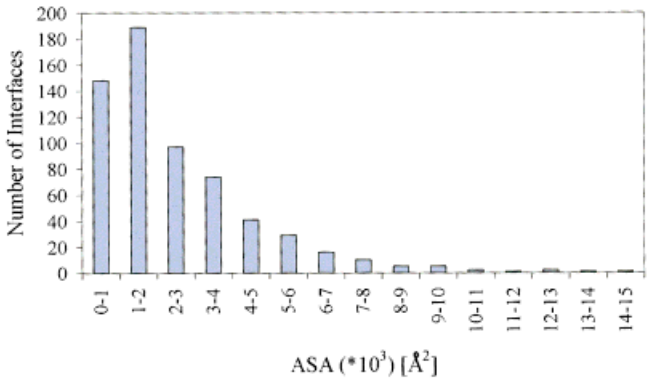


Fig. 1. Number of interfaces versus accessible surface area (ASA) in the data set of Table I.

fier (e.g., 2hhbA and 2hhbB). The family of complexes was defined as a set of complexes with homologous (i.e., belonging to the same chain family) receptors and homologous ligands. A database of representative complexes was derived from the database of all complexes. It included only one representative complex per family of complexes. The choice of the representative complex for a family of complexes was based on two criteria: the structure resolution and the surface of contact. The structure with the highest resolution often corresponded to complexes with no physical interface between components. Thus, the representative complex was designated as the structure of highest resolution with an interface surface of not less than 90% (empirical value) of the maximal interface surface in the family. The database of representative complexes (Table I) contained 621 interfaces from 440 PDB entries. Of the 621 interfaces, 404 are homodimers; i.e., the interfaces consisted of two proteins of the same family (less than 30% sequence identity).

Interface Area

The contact area of an interface between two proteins was defined as the difference in water accessible surface area (ASA) between the protein complex and the separated proteins. ASA was calculated using the program PSA,[31] and a histogram showing the distribution of interface population versus ASA is presented in Figure 1.

Residue Contacts and Frequency

A pair of amino acids i and j ($i, j = 1, 2, 3,\ldots 20$) was considered to be in contact if the distance between their C_β atoms (C_α for Gly) was smaller than a certain cutoff distance (see Results). The normalized frequency of residue i (also referred to as the amino acid composition), W_i, is defined as

$$W_i = F_i / \sum_i F_i \tag{1}$$

where F_i is the number of residues i having at least one contact with any residue across the interface. Similarly, if C_{ij} is the total number of contacts observed between residue type i and j, the normalized number of contacts between these residues is defined as $Q_{ij} = C_{ij} / \sum_{k \le l} C_{kl}$. The expected number of contacts between residue i and j is the value that would have been obtained if there were no preferences between residues of different types, i.e., $W_i \times W_j$. We estimated the likelihood of contacts between a pair of residues i and j, G_{ij}, as the log of the ratio of actual to expected number of contacts:

$$G_{ij} = A \times \log(Q_{ij}/W_i \times W_j), \tag{2}$$

where A is a constant that was arbitrarily set to 10.

This measure, used for example by Moont et al.,[27] does not take into account size differences between residues. Therefore, we normalized the number of residue–residue contacts by residue volumes, and replaced Q_{ij} with

$$Q_{ij}(v) = C_{ij} \times V_i \times V_j / \sum_{k,l} (C_{kl} \times V_k \times V_l)$$

where V_i are the residue volumes taken from Creighton.[32] Thus, the criterion for the propensity of residue–residue contacts is

$$G_{ij}(v) = A \times \log(Q_{ij}(v)/W_i \times W_j) \tag{3}$$

Error Estimates

A bootstrap procedure[33,34] was used to estimate the error in $G_{ij}(v)$. The bootstrap is a sample reuse strategy in which the statistics of interest are calculated for randomly chosen subsets of the whole set of 621 interfaces. Error estimates for the full-sample statistic can then be computed by extrapolation from the subsets to the complete data sample. An important advantage of the bootstrap for our setting is that it will automatically reflect variation at both the within interface and the between interface levels. We constructed bootstrap samples of size (s) 100, 200, 300, 400, and 500 interfaces, with 500 subsets for each sample size. For each bootstrap sample, we calculated the symmetric 20 × 20 matrix of the $G_{ij}(v, s)$ values. Then, for each sample size s, we calculated the average $[\langle G_{ij}(v, s)\rangle]$ and standard deviation $[SD_{ij}(v, s)]$. In this way, we obtained five matrices of the average pairing index ($\langle G_{ij}(v, s)\rangle$) and five matrices of the standard deviations associated with each set $[SD_{ij}(v, s)]$, where s = 100, 200, 300, 400, and 500. We extrapolated over the $SD_{ij}(v, s)$ values obtained for the small sets to estimate the $SD_{ij}(v, s)$ values in the complete set of interfaces.

RESULTS

Residue–Residue Contacts

Two residues were considered to be in contact with each other provided the distance between their C_β atoms (C_α for Gly) was below a certain cutoff. The aim of the first part of the analysis was to determine the most suitable cutoff distance. To this end, we defined the frequency density as

$$P_i(r) = F_i(r)/r^2 \tag{4}$$

where $F_i(r)$ is the frequency of residue type i at the interface, subjected to the cutoff distance r (see Materials and Methods). We calculated the frequency densities of the 20 standard amino acids; four representative examples—

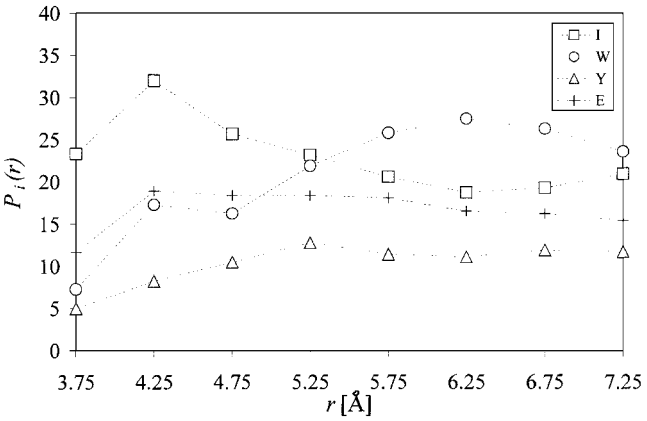


Fig. 2. Amino acid frequency density at protein–protein interfaces. r is the residue–residue distance. The $P_i(r)$ values (eq. 1) were averaged over 0.5-Å intervals of r. See text for details.

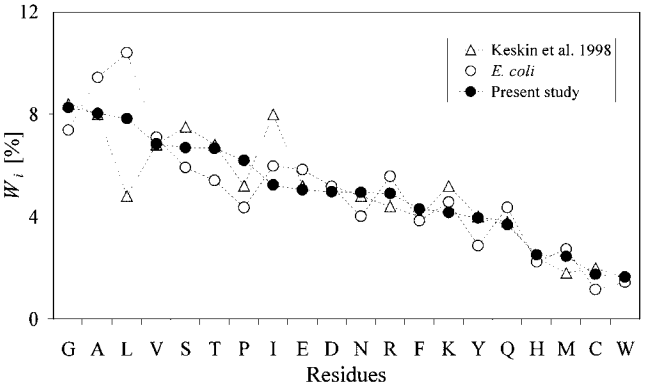


Fig. 3. Comparison of the residue composition, W_i, at protein interfaces observed in this study to that obtained in the study of Keskin et al.[25] and to the composition in the whole *Escherichia coli* genome (http://www.kazusa.or.jp/codon/). The amino acids, represented by the one-letter code, are sorted on the basis of decreasing percentage composition at the interface observed in this study.

Ile, Trp, Tyr, and Glu—are presented in Figure 2. Our results show that the maximal density of all residues, except Trp and Phe is observed before the 6–6.5-Å layer. Based on this observation we chose a cutoff distance of 6 Å for the rest of this study and refer to $F_i(0 \le r \le 6$ Å) as F_i. This choice yields an average of 83 contacts per residue pair.

Amino Acid Composition

We compared the residue composition, W_i, at protein interfaces observed in this study with that obtained by Keskin et al.[25] and with the composition in whole genomes (Fig. 3). The correlations between the residue composition in this study and in the whole *Escherichia coli,* the *Caenorhabditis elegans*, and the human immunodeficiency virus (HIV) genomes are $R = 0.91$, $R = 0.82$, and $R = 0.76$, respectively. The correlation between the residue composition in the study conducted by Keskin et al. and whole genomes is somewhat lower (e.g., $R = 0.87$ for the *E. coli* genome), presumably because we used about three times as many complexes compared with those used by Keskin's

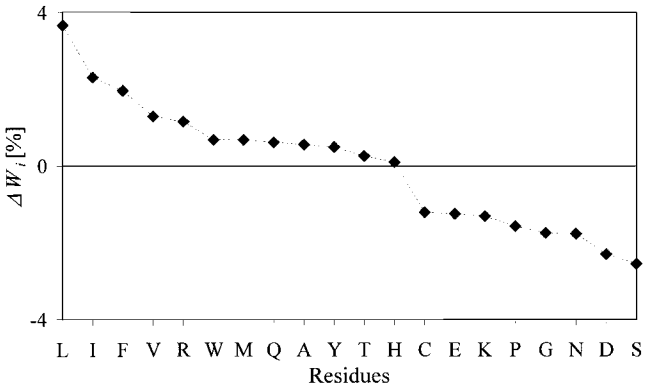


Fig. 4. The difference in residue composition between large and small interfaces. ΔW is the percentage composition at large interfaces minus that at small interfaces. The large interfaces were represented by a set of 72 complexes with contact areas of more than 5,000 Å^2, and the small interfaces were represented by a set of 148 interfaces with contact areas less than 1,000 Å^2. The residues are sorted on the basis of the difference in composition.

group. The residue composition calculated for the interprotein contacts closely resembles that calculated for the intraprotein contacts in the same set of proteins (correlation coefficient $R = 0.96$; data not shown). This is another indication of the statistical strength of the ensemble of protein–protein interfaces used in this study. The overall conclusion from the comparison of residue composition at the interprotein interface to that at intraprotein contacts and at whole genomes is that protein interfaces are not significantly different than other regions of the protein.[10]

To examine the dependence of residue composition on the size of the protein–protein interface, we calculated the percentage difference between the distributions obtained for complexes with more than 5,000 Å^2 and less than 1,000 Å^2 buried surface area (Fig. 4). Hydrophobic residues occurred more often in large contact surfaces, while polar residues prevailed in small surfaces. The exception was Arg, which was more common in large than in small contact surfaces.

Residue–Residue Contact Preferences

We calculated the number of contacts C_{ij} between residue types i and j (Table II). From the C_{ij} values, we estimated the prevalence of contacts between each pair of residues i and j using the pairwise index G_{ij} and the volume-normalized pairwise index $G_{ij}(\nu)$, as explained in the Methods. Both G_{ij} and $G_{ij}(\nu)$ were proportional to the logarithm of the fraction of actual and expected counts in the set of 621 complexes. The calculated values of G_{ij} and $G_{ij}(\nu)$ are given in the 20 × 20 symmetric matrices in Tables III and IV. For better visualization, the color-coded representations of these matrices are shown in Figure 5A and 5B.

The residues in Figure 5A are arranged according to their hydrophobicity, using the Kyte and Doolittle hydropathy index[35]; the overall tendency of hydrophobic residues to interact with each other is evident (e.g., Val–Val). In contrast, pairs of hydrophobic–hydrophilic residues were associated with low contact values (e.g., Phe–Asp). The

TABLE II. Residue–Residue Contacts, C_{ij}[†]

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	78	152	137	82	23	50	174	155	109	87	32	77	110	35	73	56	70	56	49	60
V	152	179	193	109	46	63	252	182	160	148	23	81	156	52	113	79	99	82	100	101
L	137	193	180	110	45	76	256	178	130	143	43	83	138	74	107	81	85	99	72	118
F	82	109	110	62	27	38	136	101	88	78	22	61	104	27	51	49	39	60	40	53
C	23	46	45	27	43	11	61	59	33	59	6	18	47	20	30	16	21	17	18	23
M	50	63	76	38	11	28	72	75	41	47	11	30	53	22	40	30	21	31	27	27
A	174	252	256	136	61	72	228	245	203	215	47	106	195	83	147	103	152	163	108	110
G	155	182	178	101	59	75	245	192	231	198	43	115	188	84	116	117	165	140	129	147
T	109	160	130	88	33	41	203	231	123	182	42	74	162	51	115	63	171	118	92	101
S	87	148	143	78	59	47	215	198	182	147	32	78	153	42	138	84	176	127	95	104
W	32	23	43	22	6	11	47	43	42	32	7	21	76	17	11	8	18	21	21	43
Y	77	81	83	61	18	30	106	115	74	78	21	55	91	43	66	26	41	60	52	56
P	110	156	138	104	47	53	195	188	162	153	76	91	97	51	118	89	94	129	90	102
H	35	52	74	27	20	22	83	84	51	42	17	43	51	28	30	31	69	34	22	39
E	73	113	107	51	30	40	147	116	115	138	11	66	118	30	56	42	46	79	87	103
Q	56	79	81	49	16	30	103	117	63	84	8	26	89	31	42	36	67	66	40	54
D	70	99	85	39	21	21	152	165	171	176	18	41	94	69	46	67	55	122	74	101
N	56	82	99	60	17	31	163	140	118	127	21	60	129	34	79	66	122	93	59	74
K	49	100	72	40	18	27	108	129	92	95	21	52	90	22	87	40	74	59	36	31
R	60	101	118	53	23	27	110	147	101	104	43	56	102	39	103	54	101	74	31	38

[†]The average number of contacts is 83 and the standard deviation is 55.

charged residues (Arg, Lys, Asp, Glu, and His) showed specific patterns of preferences, according to their charges. Pairs consisting of residues of the same charge (e.g., Lys–Lys, Lys–Arg, Asp–Asp, and Asp–Glu) were associated with low contact indices, while pairs of opposite-charged residues (e.g., Glu–Arg, Glu–Lys) were associated with high contact indices. Of particular interest is the seemingly high preference of small residues, such as Gly and Ala, to pair (Table III and Fig. 5A). This tendency is difficult to rationalize; the volume normalization of Figure 5B reversed it. This suggests that the apparent tendency of small residues to pair emerges from the cutoff distance of 6 Å used to define residue–residue contacts.

Contacts between pairs of hydrophobic residues produced the highest $G_{ij}(\nu)$ values, indicating that such pairing is very likely, while pairs of hydrophobic-polar residues were very unlikely (Fig. 5B). The polar residues showed intermediate $G_{ij}(\nu)$ values to pair with each other. On average, the charged residues tended to pair with each other subjected to charge complementarity, as expected. Finally, small residues were less likely to interact with each other than were large residues, presumably because the interaction strength is essentially proportional to the contact area.[30,36,37]

Specific Residue–Residue Contacts

A detailed analysis of the main peaks and troughs of Figure 5B (and Table IV) is summarized in the following discussion.

Cys–Cys

One of the highest peaks of Figure 5B was obtained for Cys–Cys pairs ($G_{CC}(\nu) = 7.65$); a closer examination of the C_β–C_β distances between these 41 pairs is presented in Figure 6. Based on their C_β–C_β distances, it is evident that the Cys–Cys pairs may be divided into two groups with distances of greater than and less than ~4.5 Å. Indeed, with two exceptions (obtained for entries 1rco and 1mka), PDB annotations regarding the presence or absence of disulfide bonds correlated well with our observations. (Examination of the 3D structures of these complexes using the InsightII software package (MSI, San Diego, CA) suggested that the Cys–Cys pair of PDB entry 1rco may actually not be an exception because the sulfide atoms are close to each other and at the right orientation for a covalent bond.) On the basis of this observation, we calculated a separate pairwise index for each of the two groups. The pairwise index value obtained for the Cys–Cys pairs that are disulfide bonded to each other was 6.37, while a value of 1.73 was obtained for the others (Table IV, box).

Hydrophobic–hydrophobic

The most common residue pairing, as reflected in the indices of Figure 5B, involved interactions between hydrophobic residues, especially the bulky aromatic residues. Four examples of this type of interaction are presented here: Trp–Leu (Fig. 7A), Trp–Tyr (Fig. 7B) and two Trp–Pro pairs (Fig. 7C,D). Most of the hydrophobic–hydrophobic pairs showed a close proximity of several aliphatic and aromatic carbon atoms although the size of the contact area between the residues varied from one case to another. In most of the cases we checked, both residues were totally, or at least partially, buried inside the core of the complex. In addition to the main hydrophobic interactions, some pairs showed a preference of aromatic rings to be oriented parallel to each other (Fig. 7B).

The Trp–Pro pairs we examined showed a broad range of angles and contact area sizes. However, it was possible to separate them roughly into two main groups. The first (Fig. 7C), showed extensive face-to-face interaction between the Trp rings and the Pro ring, which resembled the

TABLE III. Residue–Residue Contact Preferences, G_{ij}

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	−0.67	1.07	0.03	0.40	−1.21	0.68	0.95	0.33	−0.27	−1.26	0.49	0.49	0.09	−0.97	−0.80	−0.60	−0.92	−1.87	−1.70	−1.53
V	1.07	0.62	0.36	0.48	0.64	0.53	1.40	−0.13	0.24	−0.12	−2.11	−0.45	0.45	−0.41	−0.06	−0.26	−0.57	−1.37	0.23	−0.43
L	0.03	0.36	−0.53	−0.07	−0.05	0.76	0.88	−0.82	−1.25	−0.85	0.02	−0.93	−0.67	0.53	−0.89	−0.74	−1.82	−1.14	−1.78	−0.34
F	0.40	0.48	−0.07	0.04	0.34	0.35	0.74	−0.68	−0.34	−0.89	−0.29	0.33	0.70	−1.25	−1.51	−0.32	−2.61	−0.71	−1.73	−1.21
C	−1.21	0.64	−0.05	0.34	6.26	−1.13	1.16	0.90	−0.70	1.81	−2.02	−1.06	1.16	1.36	0.10	−1.28	−1.39	−2.28	−1.29	−0.93
M	0.68	0.53	0.76	0.35	−1.13	1.45	0.41	0.47	−1.23	−0.65	−0.86	−0.32	0.21	0.30	−0.13	−0.02	−2.86	−1.15	−1.01	−1.71
A	0.95	1.40	0.88	0.74	1.16	0.41	0.26	0.46	0.57	0.80	0.29	0.02	0.72	0.92	0.38	0.19	0.59	0.91	−0.13	−0.76
G	0.33	−0.13	−0.82	−0.68	0.90	0.47	0.46	−0.72	1.01	0.33	−0.21	0.25	0.44	0.85	−0.77	0.62	0.82	0.13	0.52	0.38
T	−0.27	0.24	−1.25	−0.34	−0.70	−1.23	0.57	1.01	−0.79	0.89	0.62	−0.73	0.72	−0.39	0.12	−1.13	1.91	0.32	−0.02	−0.32
S	−1.26	−0.12	−0.85	−0.89	1.81	−0.65	0.80	0.33	0.89	−0.06	−0.58	−0.52	0.46	−1.25	0.90	0.10	2.02	0.62	0.10	−0.21
W	0.49	−2.11	0.02	−0.29	−2.02	−0.86	0.29	−0.21	0.62	−0.58	−1.09	−0.12	3.51	0.92	−3.99	−4.02	−1.79	−1.10	−0.36	2.05
Y	0.49	−0.45	−0.93	0.33	−1.06	−0.32	0.02	0.25	−0.73	−0.52	−0.12	0.25	0.49	1.14	−0.02	−2.71	−2.02	−0.35	−0.23	−0.61
P	0.09	0.45	−0.67	0.70	1.16	0.21	0.72	0.44	0.72	0.46	3.51	0.49	−1.18	−0.07	0.55	0.69	−0.37	1.03	0.21	0.05
H	−0.97	−0.41	0.53	−1.25	1.36	0.30	0.92	0.85	−0.39	−1.25	0.92	1.14	−0.07	1.24	−1.49	0.02	2.20	−0.85	−2.00	−0.22
E	−0.80	−0.06	−0.89	−1.51	0.10	−0.13	0.38	−0.77	0.12	0.90	−3.99	−0.02	0.55	−1.49	−1.80	−1.69	−2.59	−0.22	0.95	0.97
Q	−0.60	−0.26	−0.74	−0.32	−1.28	−0.02	0.19	0.62	−1.13	0.10	−4.02	−2.71	0.69	0.02	−1.69	−1.00	0.41	0.36	−1.07	−0.47
D	−0.92	−0.57	−1.82	−2.60	−1.39	−2.86	0.59	0.82	1.91	2.02	−1.79	−2.02	−0.37	2.20	−2.59	0.41	−1.74	1.74	0.31	0.95
N	−1.87	−1.37	−1.14	−0.71	−2.28	−1.15	0.91	0.13	0.32	0.62	−1.10	−0.35	1.03	−0.85	−0.22	0.36	1.74	0.58	−0.66	−0.38
K	−1.70	0.23	−1.78	−1.73	−1.29	−1.01	−0.13	0.52	−0.02	0.10	−0.36	−0.23	0.21	−2.00	0.95	−1.07	0.31	−0.66	−2.06	−3.41
R	−1.53	−0.43	−0.34	−1.21	−0.93	−1.71	−0.76	0.38	−0.32	−0.21	2.05	−0.61	0.05	−0.22	0.97	−0.47	0.95	−0.38	−3.41	−3.23

interaction between two aromatic residues (Fig. 7B). The second showed a "side-to-side" orientation of the two residues, in which the Trp seemed to accommodate the Pro ring in its side (Fig. 7D). The abundance of the side-to-side orientation is probably due to the perfect geometrical fit between the two rigid residues in this orientation.

Hydrophobic-charged

Surprisingly, the most favorable residue–residue preference in our database was obtained for the Trp–Arg couple (G_{WR} (ν) = 8.57). Figure 8 shows a pair of Trp–Arg residues in a typical orientation, with the aliphatic carbon atoms of Arg and the aromatic carbon atoms of Trp in close proximity. The guanidinium group of Arg is solvent exposed, which presumably leads to further stabilization. That is, any different orientation, which involves a buried guanidinium, would be less stable than the one of Figure 8.

Oppositely charged residues

The residue pairing indices of Figure 5B indicate that oppositely charged residues tend to be in contact with each other. Figure 9A,B shows pairs of Arg–Glu in two typical orientations with respect to each other. The relative orientation of the charged groups of both residues suggests electrostatic attraction between them. This is typical; the charged groups were spatially close to each other in the vast majority of the observed pairs, suggesting that they interact with each other. The observed Arg–Glu interfacial pairs included a broad range of residue–residue side-chain distances and angles, suggesting the existence of a variety of electrostatic interactions, including salt bridges and hydrogen bonding. An example of a hydrogen bond between Arg–Glu is shown in Figure 9A. In most cases, at least one of the residue's side-chains is significantly exposed to the solvent, presumably to stabilize the protein–protein complex further. Figure 9B shows a case in which the charged groups of both residues are solvent exposed.

Interestingly, most of the interface pairs between residues of opposite charges in the data set show, in addition to the electrostatic interactions, close packing between aliphatic carbon atoms on both side-chain residues (e.g., Fig. 9B), suggesting that hydrophobic interactions may add to the pairing preferences. The extent of these hydrophobic interactions varied significantly from one pair to another. However, most of the pairs we checked had at least two aliphatic carbon atoms that were closely packed.

Residues with similar charges

The pairwise indices shown in Figure 5B indicate very low propensities for the negatively charged amino acids to pair with each other. In contrast, the positively charged residues, Arg and Lys, had an average tendency to pair among each other. Figure 10A,B show, typical orientations of Lys–Lys pairs. The orientations are such that the positively charged amine groups on each residue are as far as possible from each other, presumably to minimize the

TABLE IV. Residue–Residue Contact Preferences, Volume Normalized, $G_{ij}(v)$[†]

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	3.89	4.91	4.59	5.33	1.76	5.25	2.84	0.77	3.05	1.00	6.24	5.61	3.27	3.38	3.20	3.60	2.30	1.59	3.23	3.80
V	4.91	3.74	4.20	4.69	2.89	4.37	2.57	−0.41	2.83	1.42	2.92	3.95	2.90	3.21	3.22	3.22	1.93	1.36	4.45	4.18
L	4.59	4.20	4.03	4.86	2.93	5.32	2.77	−0.37	2.07	1.41	5.77	4.19	2.50	4.88	3.12	3.46	1.40	2.31	3.15	4.99
F	5.33	4.69	4.86	5.34	3.68	5.28	3.00	0.14	3.34	1.75	5.83	5.83	4.25	3.47	2.87	4.25	0.99	3.11	3.57	4.49
C	1.76	2.89	2.93	3.68	7.65	1.84	1.46	−0.25	1.03	2.48	2.14	2.47	2.74	4.12	2.51	1.33	0.24	−0.42	2.05	2.81
M	5.25	4.37	5.32	5.28	1.84	6.02	2.30	0.91	2.09	1.61	4.89	4.81	3.38	4.65	3.88	4.18	0.36	2.30	3.93	3.62
A	2.84	2.57	2.77	3.00	1.46	2.30	−0.52	−1.77	1.21	0.39	3.37	2.47	1.22	2.59	1.71	1.72	1.13	1.69	2.13	1.90
G	0.77	−0.41	−0.37	0.14	−0.25	0.91	−1.77	−4.40	0.21	−1.53	1.42	1.25	−0.51	1.08	−0.89	0.70	−0.08	−0.54	1.33	1.59
T	3.05	2.83	2.07	3.34	1.03	2.09	1.21	0.21	1.27	1.91	5.12	3.14	2.65	2.71	2.88	1.82	3.88	2.52	3.67	3.77
S	1.00	1.42	1.41	1.75	2.48	1.61	0.39	−1.53	1.91	−0.09	2.87	2.30	1.33	0.80	2.60	2.00	2.94	1.77	2.74	2.82
W	6.24	2.92	5.77	5.83	2.14	4.89	3.37	1.42	5.12	2.87	5.85	6.19	7.87	6.46	1.20	1.37	2.62	3.54	5.76	8.57
Y	5.61	3.95	4.19	5.83	2.47	4.81	2.47	1.25	3.14	2.30	6.19	5.93	4.22	6.05	4.54	2.05	1.76	3.66	5.26	5.28
P	3.27	2.90	2.50	4.25	2.74	3.38	1.22	−0.51	2.65	1.33	7.87	4.22	0.60	2.89	3.17	3.50	1.46	3.09	3.75	3.99
H	3.38	3.21	4.88	3.47	4.12	4.65	2.59	1.08	2.71	0.80	6.46	6.05	2.89	5.37	2.30	4.00	5.20	2.38	2.72	4.90
E	3.20	3.22	3.12	2.87	2.51	3.88	1.71	−0.89	2.88	2.60	1.20	4.54	3.17	2.30	1.65	1.95	0.08	2.68	5.32	5.75
Q	3.60	3.22	3.46	4.25	1.33	4.18	1.72	0.70	1.82	2.00	1.37	2.05	3.50	4.00	1.95	2.83	3.26	3.45	3.50	4.50
D	2.30	1.93	1.40	0.99	0.24	0.36	1.13	−0.08	3.88	2.94	2.62	1.76	1.46	5.20	0.08	3.26	0.13	3.85	3.90	4.94
N	1.59	1.36	2.31	3.11	−0.42	2.30	1.69	−0.54	2.52	1.77	3.54	3.66	3.09	2.38	2.68	3.45	3.85	2.92	3.17	3.85
K	3.23	4.45	3.15	3.57	2.05	3.93	2.13	1.33	3.67	2.74	5.76	5.26	3.75	2.72	5.32	3.50	3.90	3.17	3.24	2.29
R	3.80	4.18	4.99	4.49	2.81	3.62	1.90	1.59	3.77	2.82	8.57	5.28	3.99	4.90	5.75	4.50	4.94	3.85	2.29	2.87

CYS–CYS (DB)	6.37	CYS–CYS	1.73

[†]The box at the bottom gives $G_{ij}(v)$ values obtained for Cys–Cys pairs that are, and that are not, involved in disulfide bonds separately.

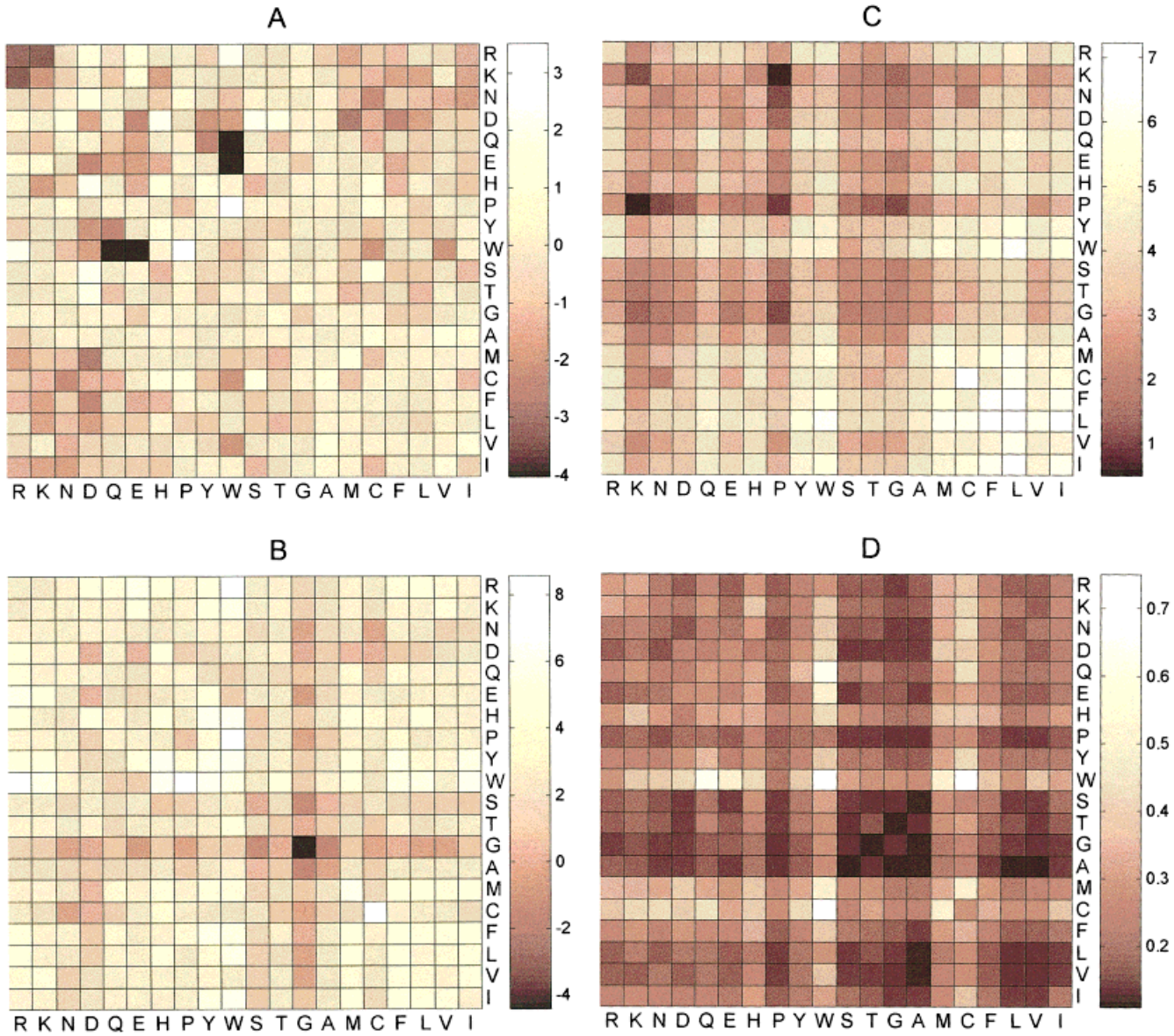


Fig. 5. Color-coded representation of the pairwise contact matrices. **A:** Pairing preferences, G_{ij}. **B:** Volume-normalized pairing preferences, $G_{ij}(\nu)$. **C:** Pairwise indices used by Keskin et al.[25] **D** Standard errors estimated for the contact indices in B. The amino acids are sorted using the Kyte and Doolittle hydropathy index.[38]

electrostatic repulsion between them. The amine groups are solvent-exposed and/or involved in hydrogen bonds to diminish the repulsion further. The most interesting observation in the figure is the extended contact area between the aliphatic chains of these residues. The close packing between the residues suggests that hydrophobic interactions are responsible for the pairing propensity observed in Figure 5B. In contrast to Arg and Lys, the negatively charged amino acids have short side-chains of very little capability for hydrophobic interactions, which is presumably why they did not portray similarly high pairing propensities with each other.

Error Estimation

We used Bickel and Yahav's[33] adaptation of the bootstrap procedure[34] for error estimation. This method is computationally intensive, so we carried it out only for the $G_{ij}(\nu)$ set of indices in Figure 5B. The procedure is based on the construction of randomly chosen subsets of the whole set of 621 interfaces and extrapolation from these subsets to the complete data set (see Materials and Methods).

The results obtained for an Arg–Arg residue pair are plotted in Figure 11 together with the $G_{RR}(\nu)$ value obtained for this pair in the whole set of 621 interfaces. A low value of $\langle G_{RR}(\nu, 100)\rangle = 2.12$ was obtained for the

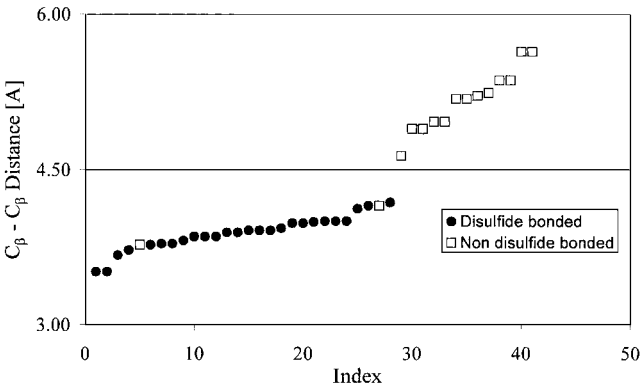


Fig. 6. The C_β–C_β distances for Cys–Cys pairs. The distances are sorted in ascending order. It is evident that the Cys–Cys pairs can be separated into two groups with distances greater than and less than ~4.5 Å. With two exceptions, a C_β–C_β distance shorter than 4.5 Å corresponds to a disulfide bond, according to the PDB annotations (filled circles).

subset of 100 interfaces. The $\langle G(\nu, s)\rangle$ value increased steadily with the number of interfaces in the subset until, at sample size 300, it plateaued at $\langle G_{RR}(\nu, 300)\rangle = 2.82$, which is very close to the value obtained for the whole set; $G_{RR}(\nu) = 2.87$.

It is evident from Figure 11 that the standard deviation decreases with subset size, as it should. We estimated the error in the $G_{RR}(\nu)$ value obtained for the Arg–Arg pair in the complete set of interfaces by extrapolation from the values of $SD_{RR}(\nu, s)$ obtained for this pair in the subsets of 300, 400, and 500 interfaces. The relation between $SD_{RR}(\nu, s)$ and the number of interfaces, N, for an Arg–Arg pair obtained from the subsets of 300, 400, and 500 interfaces is

$$\log(SD_{RR}(\nu, s)) = -0.0015s + 0.4541 \qquad (5)$$

Substituting the number of interfaces in the whole data set, 621, into this expression gives a value of $SD_{RR}(\nu) = 0.33$ for the standard error in $G_{RR}(\nu)$ obtained for an Arg–Arg pair.

We could have derived error estimates for each of the residue pairs using the procedure described above for the Arg–Arg pair. Instead, we relied on the fact that the standard deviation of an average value decreases in inverse proportion to the square root of the sample size. There are two components to this sample size: the overall frequency of contacts and the number of interfaces included in each bootstrap sample. These relationships are demonstrated in Figure 12. The three sets of points show the decrease in standard deviation as the bootstrap sample increases from 300 to 400 to 500. Within each set, the standard deviation is plotted against the log of the number of contacts and shows the expected decreasing relation. Thus a small sample of the residue pairs can be used to determine the decrease for all of the pairs when extrapolating from the bootstrap sample sizes to the actual sample size of 621 interfaces. We chose 10 residue pairs with varying numbers of contacts, including one of the pairs with the minimal number of contacts and the pair with the maximal number of contacts. We repeated the calculations of Figure 11 and fitted a linear curve similar to eq. (5) to each of these 10 pairs. The results, plotted in Figure 12, yielded the 10 extrapolated values of $SD_{ij}(\nu)$ for the complete data set (diamonds). We then fitted the linear curve ($R = 0.99$):

$$\log(SD_{ij}(\nu)) = -0.5055 \log(C_{ij}(\nu)) + 0.2686 \qquad (6)$$

to these 10 values and used the curve to calculate the standard errors of Table V. The average standard deviation in the complete data over all residue pairs is $\langle SD_{ij}(\nu)\rangle = 0.24$, which is about 8% of the average pairing index $\langle G_{ij}(\nu)\rangle = 2.91$, and the maximal standard deviation is 0.75.

Figure 5D is a color-coded representation of the error estimates in Table V. There is a strong correlation between the number of contacts per residue pair and the standard error (see Methods), which implies that residue pairs with large $G_{ij}(\nu)$ values tend to have small standard errors, while residue pairs with small $G_{ij}(\nu)$ values have larger standard errors. Comparison of Figures 5B and 5D confirms this. The hydrophobic–hydrophobic zone, from Ile to Ser (Met and Cys excluded), is associated with large pairing indices $[G_{ij}(\nu)]$ and low standard errors. The hydrophobic–hydrophilic region (the upper left or lower right corners) is typically associated with low $G_{ij}(\nu)$ values and high standard errors, and hydrophilic–hydrophilic pairs may have large or small standard errors, depending on their attractive or repulsive tendency as reflected in $G_{ij}(\nu)$. The standard error matrix (Fig. 5D) also shows very high values for rare residues, such as Trp, Cys, and Met.

For the vast majority of the residue pairs, $SD_{ij}(\nu)$ is small in magnitude compared to $G_{ij}(\nu)$, which is an indication of the statistical strength of our analysis. However, a large $|SD_{ij}(\nu)|$ to $|G_{ij}(\nu)|$ ratio was observed for some of the pairs, mainly those involving at least one small residue, such as Gly, Ala, Cys, or Ser. In a few extreme cases, e.g., Gly–Phe, Gly–Cys, Gly–Asp, Cys–Asn, Asp–Glu, Asp–Asp, and Ser–Ser, the standard error is larger in magnitude than the pairing index, which makes it impossible to tell if pairing is favorable or unfavorable.

DISCUSSION

We used a nonredundant set of 621 protein interfaces to characterize protein–protein interactions and to deduce general principles at the atomic and amino acid levels. Specifically, we estimated the residue composition (W_i) and the tendency of the 20 standard amino acids to pair with each other at the protein interface $[G_{ij}(\nu)]$.

This is a statistical survey, and the validity of the results strongly depends on the size of the interface data set. Our data set is significantly larger than data sets used in previous studies,[16,25] and we examined the statistical strength using two criteria. First, we compared the residue composition at the interface with the residue composition in whole genomes and found the distributions to be very similar. Figure 3 shows the close correlation between W_i at the interface to that of the *E. coli* genome and similarly high correlations were observed for W_i derived from other

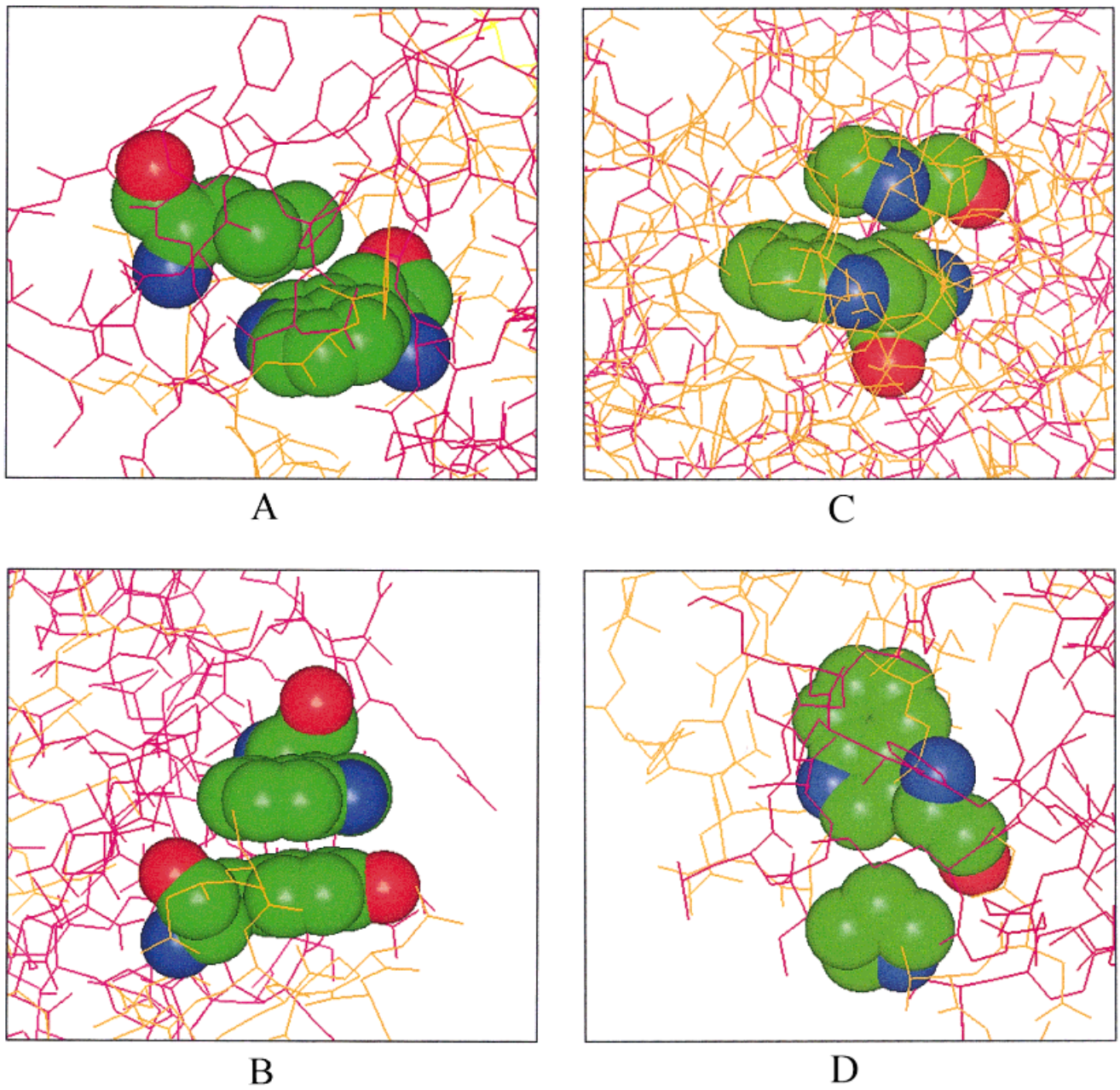


Fig. 7. Contacts between hydrophobic residues. **A:** Trp–Leu (PDB entry: 2pka; TRP-B141, LEU-A73). **B:** Trp–Tyr (1apy; TRP-A21, TYR-B308). **C:** Trp–Pro (1noy; TRP-A362, PRO-B358). **D:** Trp–Pro (1ept; TRP-B141, PRO-C152). The two interfacial residues in contact with each other are shown in space filling model. Carbon atoms are green, oxygen atoms are red, and nitrogen atoms are blue. The two proteins in the complex are marked by orange and pink bond lines.

genomes. We then estimated the errors in $G_{ij}(\nu)$ and showed that they are very small for the vast majority of the residue pairs (Table V and Fig. 5D).

A major limitation of this study is the dependence of the results on the definition of residues that are in contact with each other. The criterion used here was the C_β–C_β distance between the residues (C_α for Gly). A cutoff value of 6 Å was chosen based on close examination (Fig. 2); the pairing indices were normalized by the volume per residue. Still, as reviewed by Tsai et al.,[38] other criteria are possible, and there is no a priori reason to assume that one is better than the other. For example, Jernigan and coworkers[25] used a criterion based on counting atomic contacts that an amino acid forms. Sternberg and coworkers[16] used a method similar to ours with a C_β–C_β cutoff distance of 8 Å and without volume normalization. Reassuringly, the results of these three studies are similar despite the differences in the methodology and in the protein

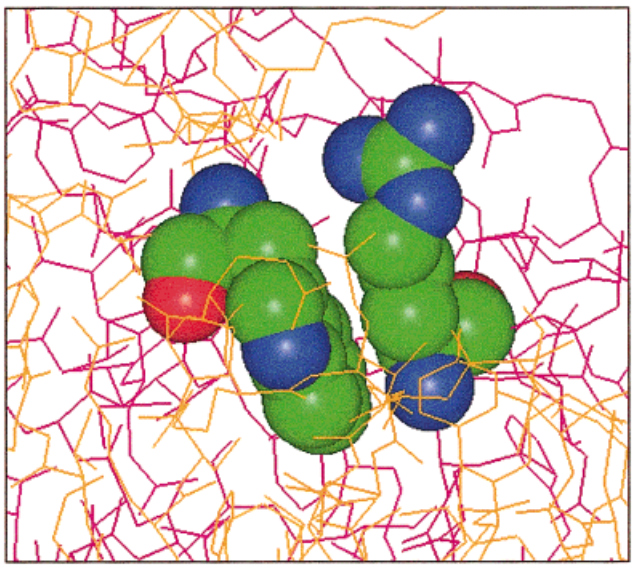


Fig. 8. Contacts between hydrophobic and charged residues pairs. The residues shown are Trp–Arg (3hhr; TRP-B169, ARG-A64). Color scheme as in Fig. 7.

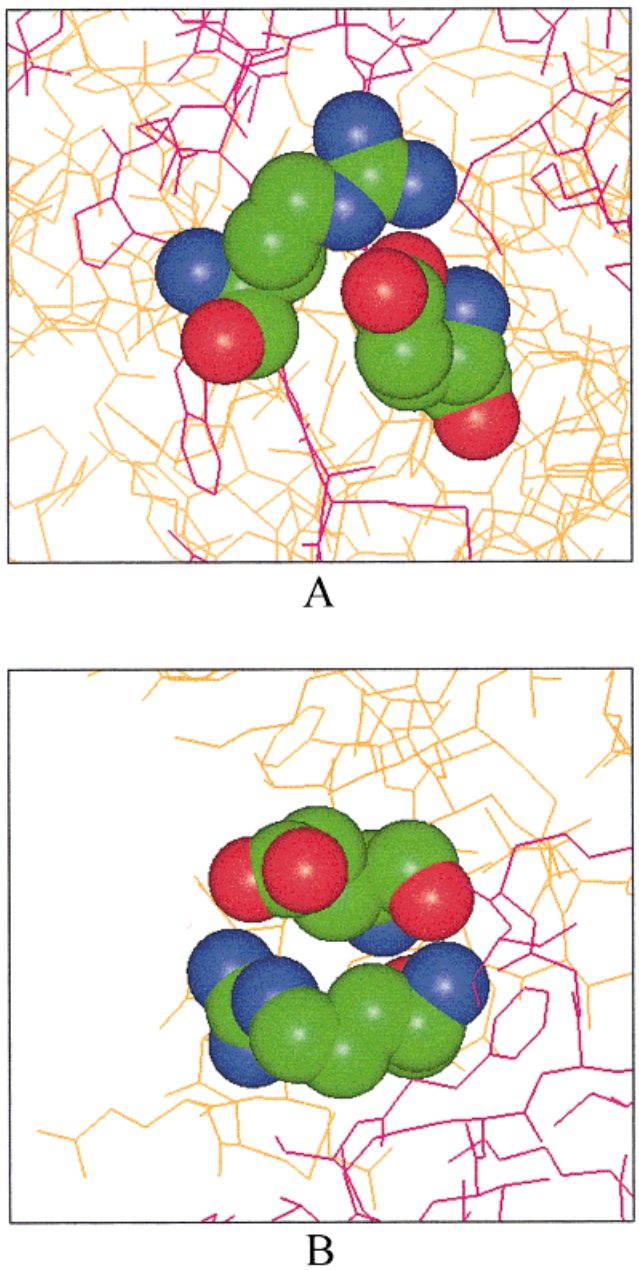


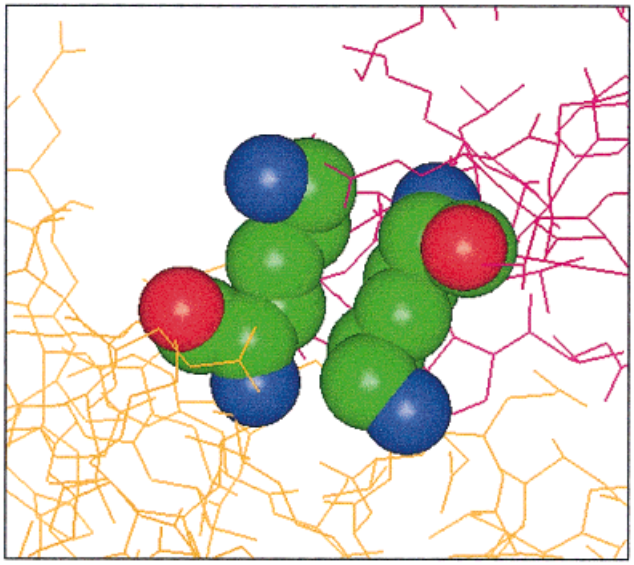
Fig. 9. Contacts between oppositely charged residues (Arg–Glu). **A:** 1osj (ARG-A144, GLU-B190). **B:** 2pol (ARG-A105, GLU-B303). Color scheme as in Fig. 7.

interface data sets. The residue composition in this study (Fig. 3) is very similar to that of Jernigan and coworkers[25]; comparison of Figure 5B and 5C shows that the values of the pairing indices in these studies are also similar (a correlation coefficient of $R = 0.52$ over 210 residue pairs). Sternberg and coworkers did not report the residue composition and their residue pairing indices were presented only graphically, which prevents a quantitative comparison. However, the overall trends observed in their study are similar to ours; contacts between hydrophobic residues are favored on average, contacts between hydrophobic and polar residues are not favored on average and contacts between charged residues follow charge complementarity.

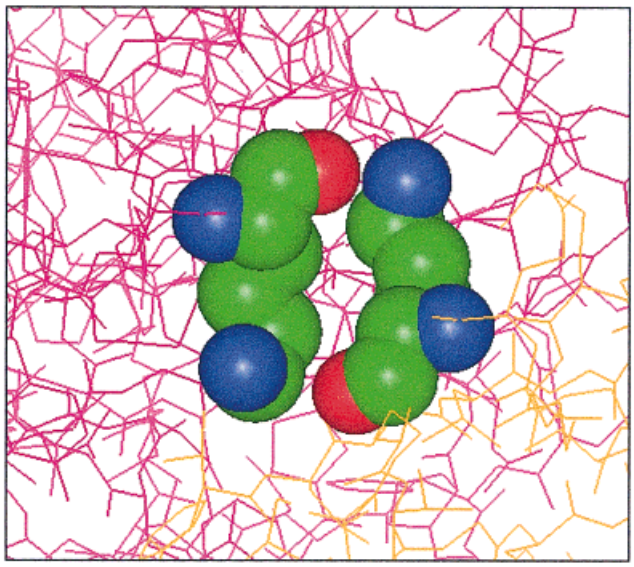
Another limitation in our study is the neglect of interfacial water molecules. These water molecules, about 18 per interface on average,[3] may be involved in hydrogen bonds and contribute to the molecular recognition between the interacting proteins. Our study is at residue rather than atomic resolution, and the neglect of water molecules is consistent with this resolution.

Our results show that the residue composition obtained at the interface (Fig. 3, filled circles) is nearly identical to the residue composition obtained for intraprotein contacts in the same protein set as expected.[25] Moreover, Figure 3 shows that the residue composition obtained at the interface is very similar to the distribution obtained in the *E. coli* genome. While the close correlation between the three distributions is an indication of the statistical strength of our results, it also implies that, on average, the protein–protein interface does not impose specific restrictions on residue identity.[3]

An important problem in studies of protein–protein interactions is the lack of credible criteria for the distinction between complexes occurring in vivo and complexes that are the artifacts of crystal packing. Complexes of proteins that interact in vivo have to be strong enough to be formed at "the biological" concentration of monomers with no help from the crystal lattice. However, experimental data on complex stability are available in only a few cases. Recently, there has some progress in assessing the free energy of complex formation.[21] However, the practical applicability of existing procedures to systematic separation of "strong" (biologically relevant) and "weak" (artifacts of crystallization) complexes is not clear. A number of empirical approaches correlate different physicochemical characteristics of the protein surfaces with the ability to form a strong complex. It has been shown that interfaces of



A



B

Fig. 10. Contacts between same charge residues. **A:** Lys–Lys (1qrd; LYS-A209, LYS-B209). **B:** Lys–Lys (1dbq; LYS-A114, LYS-B114). The color scheme is as in Fig. 7.

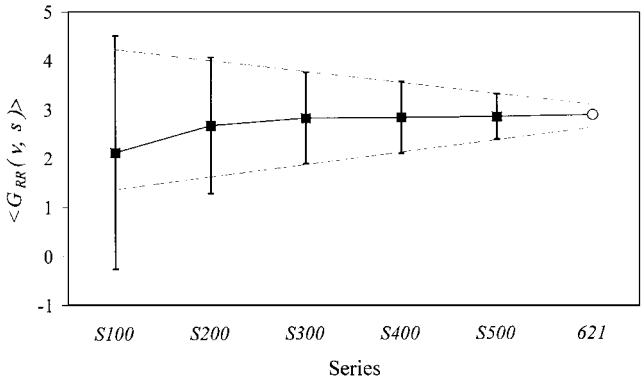


Fig. 11. Error estimate for the Arg–Arg pairing index, $G_{RR}(\nu)$. The average pairing index, $\langle G_{RR}(\nu, s)\rangle$, is plotted for five sets of bootstrap samples of protein–protein interfaces, with sample sizes of 100, 200, 300, 400, and 500 interfaces, respectively. Bars for each sample size correspond to the standard deviation from the bootstrap samples. The $G_{RR}(\nu)$ value obtained for the whole set appears as an open circle.

strongly bound proteins are generally more hydrophobic than interfaces of weakly bound proteins.[7,30] By contrast, it is also known that the contact surface of strong complexes is usually larger than that of the weak complexes.[30,36]

We used our data on the residue composition at the protein interface to correlate these two observations by comparing the hydrophobicity of large and small interfaces. We found that, on average, hydrophobic residues were more common in large interfaces and polar residues in small (Fig. 4), suggesting that biologically relevant complex formation is driven predominantly by the hydrophobic effect. That hydrophobic residues prevail in protein complexes of biological significance has been previously observed.[10] The only exception to this rule is Arg. However, Arg is a bulky residue, capable of forming hydrophobic contacts. Indeed, our survey shows that the aliphatic carbon atoms of Arg are frequently observed in close proximity to aliphatic and aromatic carbon atoms of other residues (e.g., Figs. 8, 9B). Previous analysis has shown that Arg is very common in protein interfaces,[23] and a close inspection of 23 oligomeric proteins showed that it is usually involved in hydrogen bonding at the interface.[39]

Overall, while Figure 3 shows that the protein–protein interface does not impose specific restrictions on residue identity on average, Figure 4 suggests that biologically important interfaces do.[10] For example, they are accommodative to large hydrophobic residues. Based on this observation, it seems reasonable to delete interfaces of small size from the data set when calculating the set of pairing indices $G_{ij}(\nu)$. However, the small size interfaces constitute only a small fraction of the whole data set (less than 5% of the pairwise contacts) and our calculations (data not shown) showed that the sets of G_{ij} and $G_{ij}(\nu)$ obtained from the whole interface data set (Fig. 5A,B) are essentially identical ($R \approx 0.99$) to those obtained when interfaces with a contact area of $\leq$1,000 Å^2 were removed from the calculations.

Error estimation is an important and often neglected aspect in studies, such as this one, on the determination of statistical preferences in proteins. The bootstrap procedure that we used provided an estimate of the standard errors, $SD_{ij}(\nu)$, associated with the pairing indices, $G_{ij}(\nu)$. While the $|\mathrm{SD}_{ij}(\nu)|$ to $|G_{ij}(\nu)|$ ratio was small for most of the residue pairs, in some cases $\mathrm{SD}_{ij}(\nu)$ was actually larger in magnitude than $G_{ij}(\nu)$. In these cases, typically involving contacts between small or rare residues, it is impossible to tell whether pairing is favorable. In this respect, the

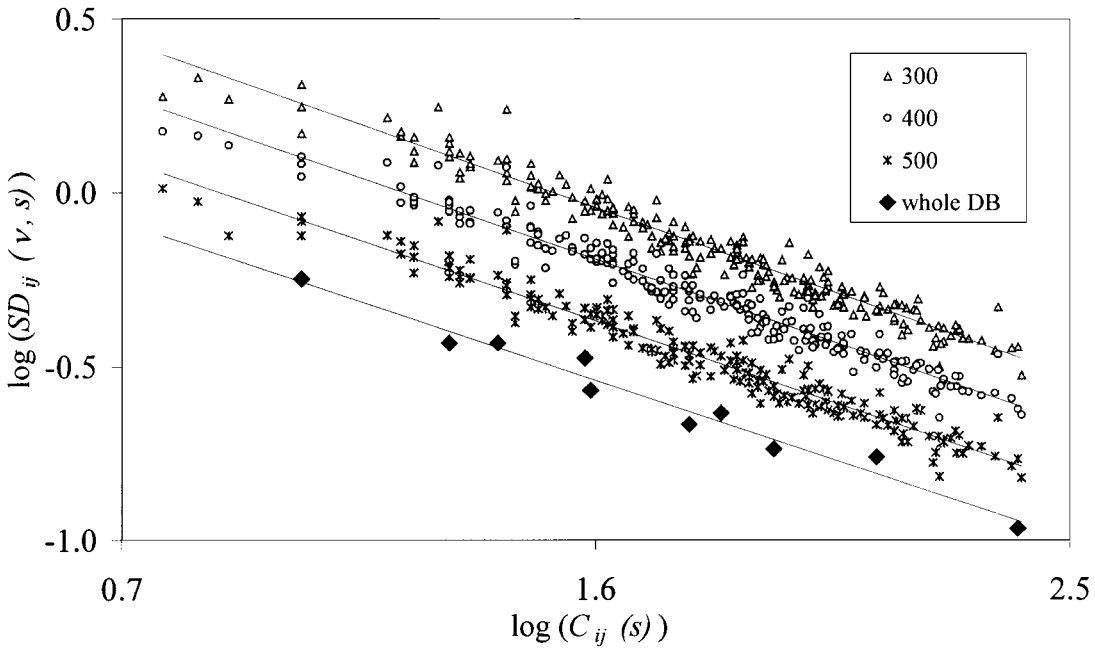


Fig. 12. Error estimates for the complete set of 621 interfaces. The log standard deviation [$SD_{ij}(\nu, s)$] is plotted vs. the log number of contacts per residue pair [$C_{ij}(s)$]. The filled diamonds represent 10 $SD_{ij}(\nu)$ values obtained for the whole set by extrapolation of the $SD_{ij}(\nu, s)$ values of the corresponding residue pairs in the three subsets. The theoretical linear relation between log [$SD_{ij}(\nu, s)$] and log [$C_{ij}(s)$] for each of the three subsets is clear. A similar relation should hold for $SD_{ij}(\nu)$ of the whole set. Thus, we fitted a linear curve to the 10 $SD_{ij}(\nu)$ values of the whole set and used the fitted curve to derive the error estimates of Table V and Fig. 5D.

TABLE V. Standard Errors, $SD_{ij}(\nu)$, for Volume-Normalized Residue–Residue Preferences

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	0.21	0.15	0.15	0.20	0.38	0.26	0.14	0.15	0.17	0.19	0.32	0.21	0.17	0.31	0.21	0.24	0.22	0.24	0.26	0.23
V	0.15	0.13	0.13	0.17	0.27	0.23	0.11	0.13	0.14	0.15	0.38	0.20	0.14	0.25	0.17	0.20	0.18	0.20	0.18	0.18
L	0.15	0.13	0.13	0.17	0.27	0.21	0.11	0.14	0.16	0.15	0.28	0.20	0.15	0.21	0.17	0.20	0.20	0.18	0.21	0.17
F	0.20	0.17	0.17	0.23	0.35	0.30	0.15	0.18	0.19	0.21	0.39	0.23	0.18	0.35	0.25	0.26	0.29	0.23	0.29	0.25
C	0.38	0.27	0.27	0.35	0.28	0.55	0.23	0.24	0.32	0.24	0.75	0.43	0.27	0.41	0.33	0.46	0.40	0.44	0.43	0.38
M	0.26	0.23	0.21	0.30	0.55	0.34	0.21	0.21	0.28	0.27	0.55	0.33	0.25	0.39	0.29	0.33	0.40	0.33	0.35	0.35
A	0.14	0.11	0.11	0.15	0.23	0.21	0.12	0.12	0.13	0.12	0.27	0.18	0.13	0.20	0.15	0.18	0.15	0.14	0.17	0.17
G	0.15	0.13	0.14	0.18	0.24	0.21	0.12	0.13	0.12	0.13	0.28	0.17	0.13	0.20	0.17	0.17	0.14	0.15	0.16	0.15
T	0.17	0.14	0.16	0.19	0.32	0.28	0.13	0.12	0.16	0.13	0.28	0.21	0.14	0.25	0.17	0.23	0.14	0.17	0.19	0.18
S	0.19	0.15	0.15	0.21	0.24	0.27	0.12	0.13	0.13	0.15	0.32	0.21	0.15	0.28	0.15	0.20	0.14	0.16	0.19	0.18
W	0.32	0.38	0.28	0.39	0.75	0.55	0.27	0.28	0.28	0.32	0.69	0.40	0.21	0.44	0.55	0.65	0.43	0.40	0.40	0.28
Y	0.21	0.20	0.20	0.23	0.43	0.33	0.18	0.17	0.21	0.21	0.40	0.24	0.19	0.28	0.22	0.36	0.28	0.23	0.25	0.24
P	0.17	0.14	0.15	0.18	0.27	0.25	0.13	0.13	0.14	0.15	0.21	0.19	0.18	0.25	0.17	0.19	0.19	0.16	0.19	0.18
H	0.31	0.25	0.21	0.35	0.41	0.39	0.20	0.20	0.25	0.28	0.44	0.28	0.25	0.34	0.33	0.33	0.22	0.31	0.39	0.29
E	0.21	0.17	0.17	0.25	0.33	0.29	0.15	0.17	0.17	0.15	0.55	0.22	0.17	0.33	0.24	0.28	0.27	0.20	0.19	0.18
Q	0.24	0.20	0.20	0.26	0.46	0.33	0.18	0.17	0.23	0.20	0.65	0.36	0.19	0.33	0.28	0.30	0.22	0.22	0.29	0.25
D	0.22	0.18	0.20	0.29	0.40	0.40	0.15	0.14	0.14	0.14	0.43	0.28	0.19	0.22	0.27	0.22	0.24	0.16	0.21	0.18
N	0.24	0.20	0.18	0.23	0.44	0.33	0.14	0.15	0.17	0.16	0.40	0.23	0.16	0.31	0.20	0.22	0.16	0.19	0.24	0.21
K	0.26	0.18	0.21	0.29	0.43	0.35	0.17	0.16	0.19	0.19	0.40	0.25	0.19	0.39	0.19	0.29	0.21	0.24	0.30	0.33
R	0.23	0.18	0.17	0.25	0.38	0.35	0.17	0.15	0.18	0.18	0.28	0.24	0.18	0.29	0.18	0.25	0.18	0.21	0.33	0.30

bootstrap procedure makes a clear distinction between the sets of $G_{ij}(\nu)$ that are meaningful and may be rationalized based on physicochemical factors, and those that are not.

Studies from many laboratories indicate the central role of the hydrophobic effect in the folding and stability of proteins.[17,35,37,40–43] The abundance of hydrophobic residues in the biologically significant protein–protein interfaces (Fig. 4) and the fact that pairs of large hydrophobic residues have the largest probability of occurrence at protein–protein interfaces (Fig. 5B) suggest that the hydrophobic effect stabilizes protein complexes as well.[19] In fact, a closer examination of pairs of polar and charged residues suggests that the hydrophobic effect derives at least partially their pairing preferences too. For example, Fig-

ures 9B, 10A, and 10B show close contacts between aliphatic carbon atoms in these charged residue pairs.

Existing docking prediction methods include approaches that concentrate on energetic considerations or on a search for the best steric fit.[16,21,44] Residue-based approaches have a long history in docking—the pioneering protein docking algorithm of Wodak and Janin[45] was based on residue approximation of protein structures. A recent study[27] showed that database-derived residue–residue preferences may be successfully used as a scoring function to suppress false-positive protein–protein matches obtained by a docking procedure. Indeed, preliminary tests that we recently carried out showed the ability of the set of pairing indices of Figure 5B in the sorting of docking decoys according to their root-mean-square deviations (RMSD) from the experimentally observed oligomeric structure (F. Glaser, A. Farkash, N. Grossaug, and N. Ben-Tal, unpublished data).

In this context, it is important to reexamine the choice of the criterion for residue–residue contacts in protein interactions. Atom-based criteria (see, e.g., Keskin et al.[25]) may reflect the physicochemical nature of interprotein interactions more accurately than those that are residue based. However, because complex formation involves adjustment of the rotameric state of the residues and rearrangement of water molecules, the C_β–C_β distance criterion used in this study may be more appropriate for docking.

ACKNOWLEDGMENTS

This work was supported by NSF grant DBI-9808093 and by South Carolina NSF EPSCoR Cooperative Agreement grant (to I.A.V.) and by grant 422-241 from the Israeli Ministry of Sports, Culture and Science (to N.B.-T.).

REFERENCES

1. Chothia C. Hydrophobic bonding and accessible surface area in proteins. Nature 1974;248:338–339.
2. Chothia C, Janin J. Principles of protein–protein recognition. Nature 1975;256:705–708.
3. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
4. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
5. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
6. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.
7. Larsen TA, Olson AJ, Goodsell DS. Morphology of protein–protein interfaces. Structure 1998;6:421–427.
8. Vakser IA, Aflalo C. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 1994;20:320–329.
9. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci 1994;3:717–729.
10. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
11. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
12. Hubbard SJ, Argos P. Cavities and packing at protein interfaces. Protein Sci 1994;3:2194–2206.
13. Janin J, Rodier F. Protein–protein interaction at crystal contacts. Proteins 1995;23:580–587.
14. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
15. Mariuzza RA, Poljak RJ. The basics of binding: mechanisms of antigen recognition and mimicry by antibodies. Curr Opin Immunol 1993;5:50–55.
16. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein–protein and protein–DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
17. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
18. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol 1997;268:570–584.
19. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
20. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
21. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
22. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
23. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
24. Dasgupta S, Iyer GH, Bryant SH, Lawrence CE, Bell JA. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. Proteins 1997;28:494–514.
25. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intramolecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
26. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
27. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
28. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A data set of protein–protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
29. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.
30. Carugo O, Argos P. Protein–protein crystal-packing contacts. Protein Sci 1997;6:2261–2263.
31. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
32. Creighton TE. Proteins: structures and molecular properties. vol. 4. New York: WA Freeman; 1996.
33. Bickel J, Yahav A. Richardson extrapolation and the bootstrap. J Am Stat Assoc 1988;83:387–393.
34. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
35. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
36. Janin J. Specific versus non-specific contacts in protein crystals. Nature Struct Biol 1997;4:973–974.
37. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
38. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Protein–protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127–152.
39. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
40. Nozaki Y, Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem 1971;246:2211–2217.
41. Kellis JT Jr, Nyberg K, Sali D, Fersht AR. Contribution of hydrophobic interactions to protein stability. Nature 1988;333:784–786.
42. Warshel A. Computer modeling of chemical reactions in enzymes and solutions. New York: John Wiley & Sons; 1991.
43. Yue K, Dill KA. Folding proteins with a simple energy function and extensive conformational searching. Protein Sci 1996;5:254–261.
44. Dixon JS. Evaluation of the CASP2 docking section. Proteins 1997;29(S1):198–204.
45. Wodak SJ, Janin J. Computer analysis of protein–protein interaction. J Mol Biol 1978;124:323–342.