SUPPLEMENTARY METHODS

Includes the following:

1. Detailed sample preparation for yeast training set and all validation sets.
2. Reference information for the top 35 physicochemical properties.
3. Computational method comparison and software versions.

Sample preparation and mass spectrometry analysis

Materials. Unless otherwise noted, chemicals were obtained from Sigma-Aldrich (St. Louis MO) and were of the highest grade available.

Yeast protein preparation. The National Cancer Institute Clinical Proteomic Technology Assessment in Cancer Program (NCI-CPTAC) prepared the yeast protein lysate by following the approach of Piening, et al., (J. Proteome Res., 5, 1527 - 1534, 2006) with modifications to allow for scale-up by a commercial concern, Boston Biochem (Cambridge, MA). Briefly, S. cerevisiae strain BY4741 (MATa, leu2D0, met15D0, ura3D0, his3D1) was grown in a 10 liter batch of rich (YPD) medium at 30 ℃ in a fermentor to an OD600 of 0.93. The yeast were harvested by continuous-flow centrifugation (yield: 5.4 g wet weight) and the cell pellet washed three times with ice cold water. The cells were lysed by incubation with ice cold trichloroacetic acid (10% final concentration in 160 ml total volume) for 1 h at 4 ℃. The protein precipitate was collected by centrifugation, washed twice with 160 mL cold 90% acetone, and pelleted again. The resulting material was lyophilized and stored at -80 ˚C. The total yield of lyophilized yeast lysate was approximately 0.75 g.

Lyophilized yeast lysate (approximately 11 mg) was reconstituted in 50 mM ammonium bicarbonate containing 2 mg/ml RapiGest SF (Waters), heated at 60 ℃ for 45 min, and sonicated for 5 min on ice. Next, 50 mM DTT in 50 mM ammonium bicarbonate was added to yield a final DTT concentration of 5 mM

and the sample incubated at 60 ℃ for 30 min. After cooling to room temperature, 200 mM iodoacetamide in water was added to yield a final concentration of 10 mM and the alkylation reaction was left to proceed at room temperature in the dark for 30 min. To quench alkylation, 100 mM DTT in 50 mM ammonium bicarbonate was added to the sample to yield a final concentration of 10 mM. Prior to the addition of trypsin, an additional volume of 50 mM ammonium bicarbonate was added to the sample to reduce the RapiGest concentration to 0.1%. Trypsin (0.5 µg/µl in 20 mM aqueous HCl) was then added to the yeast lysate sample in a 1:50 ratio to the total protein amount. The sample was digested overnight (about 18 h) at 37 ℃ with gentle swirling. After digestion, to inactivate trypsin and cleave the Rapigest, concentrated trifluoroacetic acid was added to the sample to yield a concentration of 0.5%. The sample was then incubated again at 37 ℃ for 60 min followed by centrifugation at 10,000 rpm for 10 min. The supernatant was transferred to a new sample tube and lyophilized to dryness.

Yeast mass spectrometry analysis. Aliquots of the lyophilizated yeast sample were distributed to labs at the Broad Institute (Cambridge, MA), Vanderbilt University (Nashville, TN), and New York University (New York, NY). The digested yeast sample was resuspended in 0.1% aqueous formic acid (Fluka) to yield a concentration that would correspond to approximately 60 ng/µl of total yeast protein prior to digestion. Each lab analyzed the sample using LC-MS on an LTQ Orbitrap (Thermo, San Jose, CA). The LC and mass spectrometer settings were to be followed as closely as possible by each lab. Chromatography was performed using an Agilent 1100 nanoflow chromatograph (Agilent, Palo Alto, CA) with Buffer A (0.1% formic acid) and Buffer B (90% ACN, 0.1% formic acid). A 100 um inner diameter column was pulled in-house (Sutter Instrument Model P-2000, Novato, CA) and packed with 12 cm of Jupiter (Torrance, CA) C18 resin and directly interfaced to an LTQ Orbitrap mass spectrometer. 2 µL (60 ng) of sample was injected and separated by a 120 minute gradient (~0.32%B/min.) of increasing acetonitrile from 2-40% B. MS analysis settings for

protein identification were as follows. One precursor MS scan was followed by data-dependent scans of the top 8 most abundant ions triggered in reverse order. Dynamic exclusion was enabled with a repeat count of 1, exclusion duration of 60 seconds, and an exclusion list size of 150. This process was repeated twice at NYU and Vanderbilt (note: Vanderbilt used two different LTQ Orbitrap mass spectrometers) and six times at the Broad Institute for a total of 12 runs.

Sigma48 mass spectrometry analysis. A set of 48 equimolar proteins (Universal Proteomics Standard Set, Sigma, St. Louis, MO), referred to as Sigma48, was digested using a trifluoroethanol (TFE) assisted digestion protocol[1]. The peptides were analyzed using LC-MS/MS (Agilent 1100, LTQ Orbitrap). The total run time for each replicate was 95 minutes. MS analysis settings for protein identification were as follows. One precursor MS scan was followed by data-dependent scans of the top 5 most abundant ions. Dynamic exclusion was enabled with a repeat count of 1, exclusion duration of 60 seconds and an exclusion list size of 50. This process was repeated to create four technical replicates.

HeLa_1 in-solution protein preparation. HeLa S3 cells (ATCC, Manassas VA) were cultured in RPMI 1640 medium (Invitrogen, Auckland, New Zealand) supplemented with 5% fetal bovine serum, 100 U/ml each of penicillin and streptomycin (Invitrogen), and 292 µg/ml L-glutamine (Invitrogen). Cultures were incubated at 37℃ in 5% CO2. To prepare lysates, 2.2 x 107 cells were pelleted by centrifugation at 1000 rpm and 4℃, washed with 1X PBS (Invitrogen), and centrifuged again. The resultant cell pellet was resuspended in 1 mL ice-cold modified RIPA buffer containing 50 mM Tris-HCl pH 7.8, 150 mM NaCl, 0.1% sodium deoxycholate, 1 mM EDTA, and 1% NP-40. The cell lysate was centrifuged at 14,000 rpm 4℃ for 10 min to pellet cellular membranes and nuclei. The amount of protein present in the cleared lysate was quantified by Bradford assay (Pierce a division of Thermo Fisher Scientific, Rockford, IL).

The lysate was incubated for 30 minutes at room temperature in the presence of 6M urea and 10 mM DTT to denature and reduce the proteins. Iodoacetamide was added to a final concentration of 30 mM and the lysate was incubated for an additional 30 minutes at room temperature. Additional DTT was added to neutralize excess iodoacetamide and the reaction was incubated for an additional 30 minutes at room temperature. 1 mg of protein was incubated in the presence of 50 mM Tris pH 8 and trypsin at a final concentration of 143 µg/ml with shaking overnight at 37°C. The digest was acidified with HCl and run over an Oasis MCX Extraction Cartridge (Waters, Milford, MA) (1 ml MeOH, 2 x 1 ml H2O, Sample diluted to 1 ml with water, 1 ml 0.1M HCL, elute 1 ml 50% MeOH / 5% NH4OH, speed vac to dryness) and an Oasis HLB Extraction Cartridge (Waters, Milford, MA) (resuspend in 1 ml 0.1% formic acid (Fluka) and vortex 10 min, column: 1 ml ACN, 1 ml 0.1% FA, another 1 ml FA, sample, 2 x 1 ml 0.1% FA, elute in 1 ml 80% ACN in water, freeze, speed vac to dryness).

HeLa_1 in-solution mass spectrometry analysis. The peptides were resuspended in 100 µL of 5% acetonitrile and 5% formic acid (Fluka) and vortexed before LC-MS/MS analysis (Agilent 1100, LTQ-Oribtrap). 1 µL of the peptide mixture was injected and separated by a 38 min gradient (~1.11%B/min.) of increasing acetonitrile from 7.5-50%B. MS analysis settings for protein identification were as follows. One precursor MS scan was followed by data-dependent scans of the top 10 most abundant ions triggered in reverse order. Dynamic exclusion was enabled with a repeat count of 2, exclusion duration of 60 seconds and an exclusion list size of 500. This process was repeated to create four technical replicates.

HeLa_2 GeLC protein preparation. HeLa S3 cells were cultured as described above, and lysates prepared in the same manner. Cysteines were reduced in 10 mM DTT and subsequently alkylated with five fold excess of iodoacetamide. 4x LDS buffer (Invitrogen) was added and samples were prepared for gel separation as per manufacturer's instructions. 100 µg of protein was loaded in each lane of

a 4-12% Bis-Tris gel (1.5 mm thick Nupage Invitrogen). Gels were stained with SimplyBlue Colloidal Coomassie stain (Invitrogen) and each gel lane was cut into six pieces for GeLCMS anaysis (can cite some ref). Briefly, gel pieces were cut into 1 mm square pieces and destained with ten gel volumes of 50mM ammonium bicarbonate and ethanol (50:50 v/v). Destained gel pieces were dehydrated with 100% ethanol until opaque. After aspiration of ethanol, one gel volume of 12.5 ng/µL trypsin (Promega) was added to gel pieces and allowed to rehydrate for 15 min. 50 mM ammonium bicarbonate solution was then added to the tubes until gel pieces were submerged. Digests were incubated in a mixer overnight at 37 deg C. Enzymatic digests were stopped by adding 10 µL of 1% TFA and digest supernatants set aside. A further extraction of gel pieces was performed with 300 µL of 0.1% TFA. The two peptide extracts were pooled and subjected to C18 StageTip sample clean-up as described[2]. Peptides were eluted with 50 µL of 80% acetonitrile, 0.1% TFA in water and dried in a Speedvac centrifuge to remove organic solvent.

HeLa_2 GeLC mass spectrometry analysis. The peptides were resuspended in 7 µL of 0.1% TFA and tubes vortexed before LC-MS/MS analysis (Agilent 1100, LTQ-Orbitrap). 5 µL of each peptide mixture was injected and separated by a 45 min gradient (~0.75% B/min) of increasing acetonitrile from 12-45% B. MS analysis settings for protein identification were as follows. One precursor MS scan was followed by data-dependent scans of the top 5 most abundant ions. Dynamic exclusion was enabled with a repeat count of 1, exclusion duration of 30 seconds and an exclusion list size of 500. This process was repeated to create four technical replicates.

Pull-down protein preparation and mass spectrometry analysis. Proteins from an affinity pulldown using SILAC labeled cells[3] were processed for GeLCMS analysis as described above in 'Hela_2 GelC preparation.' The entire gel lane was divided into six slices and processed for LC-MS/MS analysis as described above. LTQ-Orbitrap .raw files were extracted with using extractMSn.exe

(Thermo, Bremen Germany) and MS/MS peak lists were combined into a single Mascot generic file (.mgf) using DTAsupercharge v1.19 (http://msquant.sourceforge.net) from the CEBI group (University of Southern Denmark, Odense). A database search using Mascot v2.1.03 (MatrixScience, London) was performed with the IPIhuman database v3.32, with arginine-$^{13}C_6$, lysine-$^{13}C_6{}^{15}N_2$, and oxidized methionines as variable modifications and carbamidomethylated cysteines as a fixed mod. The precursor mass tolerance used for the search was 15 ppm and the product mass tolerance was set to 0.7 Da. The Mascot result file was parsed and peptides quantified using MSQuant v.1.4.2 (CEBI, http://msquant.sourceforge.net). Total ion intensities were obtained by summing XICs for all 'light' and 'heavy' peptide pairs for a particular protein.

Plasma protein depletion. A 200 µL of pooled of female plasma sample was depleted using the MARS Hu-14 immunodepletion columns (Agilent, Santa Clara, CA) following the manufacturer's instructions. The plasma sample was diluted 1:4 prior to injection according to the column capacity specified by the manufacturer. Immunoaffinity chromatography was performed on an Agilent 1100 pump and the collected flow through fraction of immuno depleted material was concentrated using an Amicon 3kDa MWCO. The total protein concentration, approximately 450 µg, was estimated by Bradford assay (Pierce).

Plasma protein digestion. The depleted and non-depleted plasma sample was incubated for 30 minutes at 37 ℃ temperature in the presence of 6M urea and 20 mM DTT to denature and reduce the proteins. Iodoacetamide was added to a final concentration of 50 mM and the lysate was incubated for an additional 30 minutes at room temperature in the dark. Urea was diluted to 0.6 M and the pH was adjusted to 8.0 with 1M Trizma base. Trypsin was added to achieve a 1:50 enzyme to substrate ratio and incubated overnight at 37 ℃. The digest was acidified with 1% formic acid. The sample was desalted using an Oasis HLB 1cc (30mg) Extraction Cartridge (Waters, Milford, MA) conditioned with 3 x 500 µl

Acetonitrile, followed by 4 x 500 µl 0.1% formic acid. Sample in 1 ml of 1% formic acid was loaded onto cartridge and washed with 3 x 500 µl 0.1% formic acid. Desalted peptides were eluted with 2 x 500 µl of 80% acetonitrile/0.1% formic acid frozen at -80 ℃ prior to speed vac to dryness.

Plasma peptide fractionation. Digested plasma samples (100 µg total protein) were reconstituted in 100 µL in 25% acetonitrile, pH 3.0 and fractionated by strong cation exchange (SCX) chromatography on a BioBasic 1 x 250 mm column (ThermoFisher). Separations were performed on an Agilent 1100 analytical LC system (Agilent Technologies) at a flow rate of 50 µL/min and mobile phase that consisted of 25% acetonitrile, pH 3.0 (A) and 250 mM ammonium formate / 25% acetonitrile, pH 3.0 (B). After loading 100 µL of sample onto the column, the mobile phase was held at 1% B for 15 minutes. Peptides were then separated with a linear gradient of 1-16% B in 30 minutes, 16-36% B in 27 minutes, 36-60% B in 1.3 minutes, 60-100% B in 2 minutes, held at 100% B for 10 minutes, then 1% B in 1 minute and held for 34 minutes. Twelve fractions were collected equally spaced between 0-80 minutes. Fractions 1 and 2 were pooled to create 11 fractions. Fractions were desalted using Oasis HLB 1cc (10mg) reversed phase cartridges, and eluates were dried to dryness via vacuum centrifugation. For peptides that elute early by SCX (e.g fractions 1-2), the flowthrough from the SCX separation was also desalted and analyzed by LC-MRM/MS.

Plasma mass spectrometry analysis. The peptides were resuspended in 100 µL of 3% acetonitrile and 0.5% formic acid and vortexed before LC-MS/MS analysis (Agilent 1100, LTQ-Oribtrap). A 1:3 dilution of the resuspended material was used and 1 µL of the peptide mixture was injected and separated by an 100 min gradient (~0.7%B/min.) of increasing acetonitrile from 5 -60%B. MS analysis settings for protein identification were as follows. One precursor MS scan was followed by data-dependent scans of the top 8 most abundant ions. Dynamic

exclusion was enabled with a repeat count of 2, exclusion duration of 20 seconds and an exclusion list size of 500.

Protein and peptide identification

Spectrum Mill v3.4 beta. The LTQ-Orbitrap .raw files (except HeLa GeLC and Eval-19) were searched using Spectrum Mill. The peptide identification settings were as follows: precursor mass of 0.05 Da and fragment mass of 0.7 Da, allowing up to two missed cleavages and specifying the following modifications: carbamidomethylation, carboxymethylation, carbamylated lysine, oxidized methionine and pyroglutamic acid. The data was autovalidated at the protein level with a protein score of 25 and at the peptide level using the default values. The yeast data set was searched against the SGD yeast verified database (05/11/07). The Sigma48 data set was searched against a SwissProt database containing only the 48 proteins. The HeLa In-solution and plasma data sets were searched against the human IPI database v3.35.

Mascot. LTQ-Orbitrap .raw files for HeLa GeLC and HeLa In-solution were extracted using extractMSn.exe (Thermo, Bremen Germany) and MS/MS peak lists were combined into a single Mascot generic file (.mgf) using DTAsupercharge v1.19 (http://msquant.sourceforge.net) from the CEBI group (University of Southern Denmark, Odense). A database search using Mascot v2.1.03 (MatrixScience, London) was performed with the IPIhuman database v3.32 with oxidized methionines as a variable modification and carbamidomethylated cysteines as a fixed modification. The precursor mass tolerance used for the search was 15 ppm and the product mass tolerance was set to 0.7 Da. The Mascot result file was parsed and peptides quantified using MSQuant v.1.4.2 (CEBI, http://msquant.sourceforge.net).

Reference Information for the top 35 features

The details are from the AAIndex[5].

H: AAIndex ID

D: Description

R: Reference

A: Author list

T: Title

J: Journal

Abstract: listed if present in an electronic form.

1. Mass (Matlab bioinformatics toolbox)

2. Length

3. Positive charge (Fauchere, 1988)

   H FAUJ880111

   D Positive charge (Fauchere et al., 1988)

   R LIT:1414114 PMID:3209351

   A Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V.

   T Amino acid side chain parameters for correlation studies in biology and pharmacology

   J Int. J. Peptide Protein Res. 32, 269-278 (1988)

   Abstract: Fifteen physicochemical descriptors of side chains of the 20 natural and of 26 non-coded amino acids are compiled and simple methods for their evaluation described. The relevance of these parameters to account for hydrophobic, steric, and electric properties of the side chains is assessed and their intercorrelation analyzed. It is shown that three principal components, one steric, one bulk, and one electric

(electronic), account for 66% of the total variance in the available set. These parameters may prove to be useful for correlation studies in series of bioactive peptide analogues.

4. Gas phase basicity (Zhang, 2004)[6]

Abstract: A kinetic model, based on the "mobile proton" model of peptide fragmentation, was developed to quantitatively simulate the low-energy collision-induced dissociation (CID) spectra of peptides dissociated in a quadrupole ion trap mass spectrometer. The model includes most fragmentation pathways described in the literature, plus some additional pathways based on the author's observations. The model was trained by optimizing parameters within the model for predictions of CID spectra of known peptides. A best set of parameters was optimized to obtain best match between the simulated spectra and the experimental spectra in a training data set. The performance of the mathematical model and the associated optimized parameter set used in the CID spectra simulation was evaluated by generating predictions for a large number of known peptides, which were not included in the training data set. It was shown that the model is able to predict peptide CID spectra with reasonable accuracy in fragment ion intensities for both singly and doubly charged peptide parent ions up to 2000 u in mass. The optimized parameter set was evaluated to gain insight into the collision-induced peptide fragmentation process.

5. Hydrophobicity index 3.0 pH (Cowan, 1990)

   H COWR900101

   D Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)

   R PMID:2134053

   A Cowan, R. and Whittaker, R.G.

   T Hydrophobicity indices for amino acid residues as determined by

high-performance liquid chromatography
J Peptide Res. 3, 75-80 (1990)

Abstract: The retention times of compounds on reversed-phase high-performance liquid chromatography columns are determined by their overall hydrophobicity. This paper exploits this relationship to derive hydrophobicity indices for amino acid residues. Retention times of 20 Z-amino acids and their methyl, ethyl and benzyl esters were determined under standard conditions on reversed-phase high-performance liquid chromatography at pH 3.0 and pH 7.5. Retention times of derivatives of six of the amino acids were used to calculate amino acid residue hydrophobicity values relative to glycine and leucine by a computer-based iterative procedure that used splines to optimize the smoothness of fit. Using these derived curves, values for each derivative of the remaining 14 Z-amino acids were determined and averages calculated. The curves, generated independently for pH 3.0 and pH 7.5 were effectively identical and the determined hydrophobicity values (other than for charged residues) were also similar at the two pHs. The values obtained vary significantly from other published values. Comparisons with some of the more commonly used hydrophobicity/hydropathy tables are presented. The highest correlation was found with constants determined for water/octanol partitioning of N- and C-terminal protected amino acids.

6. Fraction of site occupied by water (Krigbaum, 1979)
H KRIW790102
D Fraction of site occupied by water (Krigbaum-Komoriya, 1979)
R LIT:0502056 PMID:760806
A Krigbaum, W.R. and Komoriya, A.
T Local interactions as a structure determinant for protein molecules: II
J Biochim. Biophys. Acta 576, 204-228 (1979)

Abstract: Van der Waals interactions between sidechains are indicated to be important in determining the native state of the proteins of known structure by the following observations: 1. the average radial distribution of polarity increases continuously from the center of the molecule to its periphery. 2. nonpolar sidechains tend to occur in clusters. 3. the frequencies of long-range nearest-neighbor pairs are markedly non-random; each type of sidechain seeks nearest-neighbors of similar polarity. To investigate how these interactions affect the overall structure of the protein molecule, three simplified models are treated: a sheath-core model composed of independent residues, a modification accounting approximately for the connected nature of the chain, and a model consisting of three concentric spherical phases.

7. Hydrophobicity coefficient in RP-HPLC C18 (Wilce, 1995)

H WILM950103
D Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
R
A Wilce, M.C., Aguilar, M.I. and Hearn, M.T.
T Physicochemical basis of amino acid hydrophobicity scales: evaluation of four
new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of
peptides
J Anal Chem. 67, 1210-1219 (1995)
Abstract:

The physicochemical relationships of four new scales of amino acid hydrophobicity coefficients, derived from the reversed-phase high-performance liquid chromatographic (RP-HPLC) retention data of 1738 peptides, have been compared with 12 previously published scales of amino acid hydrophobicities. Four different reversed-phase chromatographic systems were used to obtain the experimental data, including three well-characterized hydrophobic adsorbents of different *n*-alkyl ligand chain length (*n*-octadecyl (C18), *n*-octyl (C8), and *n*-butyl (C4) ligands) and two different aquo–organic eluents, namely a binary water–acetonitrile and a ternary water–acetonitrile–2-propanol elution system, both containing 0.1% trifluoroacetic acid. Significant correlations were observed between the amino acid hydrophobicity scales derived from the RP-HPLC data of these peptides separated with the C18- or C8-silica adsorbents and the water–acetonitrile elution system and the hydrophobicity scales based on the transfer free energies of an amino acid entity from a polar to a nonpolar solvent or the statistical degree of exposure to the surrounding solvent of an amino acid side chain within a protein structure. Conversely, lower levels of correlation were observed between the previously calculated hydrophobicity scales for the common α-amino acids and the amino acid hydrophobicity coefficient values derived from this peptide structure–retention data base with the C4 ligand or the C18 ligand in combination with the water–acetonitrile–2-propanol solvent system. The results confirm that the relative ranking and magnitude of the hydrophobicities of amino acid side chains within polypeptides are dependent upon the chemical microenvironment of the interface established between the solute, the solvent, and the immobilized hydrocarbonaceous ligand. The availability of this extensive data base of peptide structure–chromatographic retention behavior has also permitted differences in peptide–nonpolar ligand interactions to be quantified in physicochemical terms, including the propensity of amino acid side chains in peptide structures to partition into different hydrophobic environments from aqueous solutions of different hydrogen-bonding or dipolar characteristics.

8. Transfer energy organic solvent water (Nozaki, 1971)

   H NOZY710101

   D Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)

R PMID:5555568

A Nozaki, Y. and Tanford, C.

T The solubility of amino acids and two glycine peptides in aqueous ethanol and
dioxane solutions

J J. Biol. Chem. 246, 2211-2217 (1971) Missing values filled with zeros

Abstract:

The solubilities of amino acids, diglycine, and triglycine have been measured in water and aqueous ethanol as well as dioxane solutions. Free energies of transfer of amino acid side chains and backbone peptide units from water to ethanol and dioxane solutions have been calculated from these data. The results show the similarity between the effects of ethanol and dioxane on the stability of those side chains and peptide units. In particular, the free energies of transfer of hydrophobic side chains to 100% ethanol and dioxane are essentially identical, and have been used to establish a hydrophobicity scale for hydrophobic side chains.

9. Partition coefficient (Garel, 1973)

H GARJ730101

D Partition coefficient (Garel et al., 1973)

R LIT:2004092b PMID:4700470

A Garel, J.P., Filliol, D. and Mandel, P.

T Coefficients de partage d'aminoacides, nucleobases, nucleosides et nucleotides dans un systeme solvant salin

J J. Chromatogr. 78, 381-391 (1973)

10. Energy transfer from out to in (95% buried) (Radzicka, 1988)

H RADA880107

D Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)

R LIT:1405051b

A Radzicka, A. and Wolfenden, R.

T Comparing the polarities of the amino acids: Side-chain distribution
coefficients between the vapor phase, cyclohexane, 1-octanol, and
neutral
aqueous solution
J Biochemistry 27, 1664-1670 (1988) (Pro missing)

11. Partition energy (Guy, 1985)

H GUYH850101
D Partition energy (Guy, 1985)
R LIT:2004051b PMID:3978191
A Guy, H.R.
T Amino acid side-chain partition energies and distribution of residues in
soluble proteins
J Biophys. J. 47, 61-70 (1985)

Abstract: Energies required to transfer amino acid side chains from water to less polar environments were calculated from results of several studies and compared with several statistical analyses of residue distributions in soluble proteins. An analysis that divides proteins into layers parallel with their surfaces is more informative than those that simply classify residues as exposed or buried. Most residues appear to be distributed as a function of the distance from the protein-water interface in a manner consistent with partition energies calculated from partitioning of amino acids between water and octanol phases and from solubilities of amino acids in water, ethanol, and methanol. Lys, Arg, Tyr, and Trp residues tend to concentrate near the water-protein interface where their apolar side-chain components are more buried than their polar side-chain components. Residue distributions calculated in this manner do not correlate well with side-chain solvation energies calculated from vapor pressures of side-chain analogs over a water phase. Results of statistical studies that classify residues as exposed to solvent or buried inside the protein interior

appear to depend on the method used to classify residues. Data from some of these studies correlate better with solvation energies, but other data correlate better with partition energies. Most other statistical methods that have been used to evaluate effects of water on residue distributions yield results that correlate better with partition energies than with solvation energies.

12. Retention coefficient at pH 2 (Guo, 1986)

H GUOD860101
D Retention coefficient at pH 2 (Guo et al., 1986)
R
A Guo, D., Mant, C.T., Taneja, A.K., Parker, J.M. and Hodges, R.S.
T Prediction of peptide retention times in reversed-phase high-performance
liquid chromatography; I. determination of retention coefficients of amino acid residues of model synthetic peptides
J J Chromatogr. 359, 499-517 (1986)

13. RF value in high salt chromatography (Weber, 1978)

H WEBA780101
D RF value in high salt chromatography (Weber-Lacey, 1978)
R LIT:2004106b PMID:691071
A Weber, A.L. and Lacey, J.C.,Jr.
T Genetic code correlations: Amino acids and their anticodon nucleotides
J J. Mol. Evol. 11, 199-210 (1978)

Abstract: The data here show direct correlations between both the hydrophobicity and the hydrophilicity of the homocodonic amino acids and their anticodon nucleotides. While the differences between properties of uracil and cytosine derivatives are small, further data show that uracil has an affinity for charged species. Although these data suggest that

molecular relationships between amino acids and anticodons were responsible for the origin of the code, it is not clear what the mechanism of the origin might have been.

14. The stability scale from knowledge based atom atom potential (Zhou, 2004)
    H ZHOH040101
    D The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004)
    R PMID:14696193
    A Zhou, H. and Zhou, Y.
    T Quantifying the effect of burial of amino acid residues on protein stability
    J Proteins 54, 315-322 (2004)

    Abstract: The average contribution of individual residue to folding stability and its dependence on buried accessible surface area (ASA) are obtained by two different approaches. One is based on experimental mutation data, and the other uses a new knowledge-based atom-atom potential of mean force. We show that the contribution of a residue has a significant correlation with buried ASA and the regression slopes of 20 amino acid residues (called the buriability) are all positive (pro-burial). The buriability parameter provides a quantitative measure of the driving force for the burial of a residue. The large buriability gap observed between hydrophobic and hydrophilic residues is responsible for the burial of hydrophobic residues in soluble proteins. Possible factors that contribute to the buriability gap are discussed. Copyright 2003 Wiley-Liss, Inc.

15. Apparent partition energies calculated from Chothia index (Guy, 1985)
    H GUYH850105
    D Apparent partition energies calculated from Chothia index (Guy, 1985)
    R PMID:3978191
    A Guy, H.R.

T Amino acid side-chain partition energies and distribution of residues in soluble proteins
J Biophys. J. 47, 61-70 (1985)

Abstract: Energies required to transfer amino acid side chains from water to less polar environments were calculated from results of several studies and compared with several statistical analyses of residue distributions in soluble proteins. An analysis that divides proteins into layers parallel with their surfaces is more informative than those that simply classify residues as exposed or buried. Most residues appear to be distributed as a function of the distance from the protein-water interface in a manner consistent with partition energies calculated from partitioning of amino acids between water and octanol phases and from solubilities of amino acids in water, ethanol, and methanol. Lys, Arg, Tyr, and Trp residues tend to concentrate near the water-protein interface where their apolar side-chain components are more buried than their polar side-chain components. Residue distributions calculated in this manner do not correlate well with side-chain solvation energies calculated from vapor pressures of side-chain analogs over a water phase. Results of statistical studies that classify residues as exposed to solvent or buried inside the protein interior appear to depend on the method used to classify residues. Data from some of these studies correlate better with solvation energies, but other data correlate better with partition energies. Most other statistical methods that have been used to evaluate effects of water on residue distributions yield results that correlate better with partition energies than with solvation energies.

16. Average relative fractional occurence in E0i (Rackovsky, 1982)
H RACS820105
D Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)

R LIT:0903736
A Rackovsky, S. and Scheraga, H.A.
T Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
J Macromolecules 15, 1340-1346 (1982)

17. Isoelectric point (Matlab bioinformatics toolbox)

18. Transfer free energy from vap to chx (Radzicka, 1988)
H RADA880103
D Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)
R LIT:1405051b
A Radzicka, A. and Wolfenden, R.
T Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral
aqueous solution
J Biochemistry 27, 1664-1670 (1988) (Pro missing)

19. Propensity of amino acids with pi helices (Fodje, 2002)
H FODM020101
D Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)
R PMID:12034854
A Fodje, M.N. and Al-Karadaghi, S.
T Occurrence, conformational features and amino acid propensities for the pi-helix
J Protein Eng. 15, 353-358 (2002)

Abstract: The most abundant helix type in proteins is the alpha-helix, accounting for about 31% of amino acid secondary structure states, while the 3(10)-helix accounts for about 4%. The pi-helix appears to be

extremely rare and is considered to be unstable. Existing secondary structure definition methods find very few within the Protein Data Bank. Using an improved pi-helix definition algorithm to search a non-redundant subset of high-resolution and well-refined protein structures, we found that almost every tenth protein contained a pi-helix. This enabled us to show for the first time that the pi-helix has structural parameters that are different from the hypothesized model values. It also has distinctive amino acid preferences and it is conserved within functionally related proteins. Features that may contribute to the stability of the pi-helical structure have also been identified. In addition to hydrogen bonds, several other factors contribute to the stability of pi-helices. The pi-helix may have some functional advantages over other helical structures. Thus, we describe cases where the side chains of functionally important residues at every fourth position within a pi-helix could be aligned and brought close together in a way that would not be allowed by any other helix type.

20. Propensity to be buried inside (Wertz, 1978)

H WERD780101

D Propensity to be buried inside (Wertz-Scheraga, 1978)

R LIT:0405105 PMID:621952

A Wertz, D.H. and Scheraga, H.A.

T Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule

J Macromolecules 11, 9-15 (1978) Adjusted values

Abstract: The x-ray structures of 20 proteins have been examined and each of the residues in these proteins was assigned to the inside or outside of the molecules and to a conformational state. The data obtained confirm that polar groups are generally found on the outside of proteins

and nonpolar residues are generally found on the inside. Seven of the amino acids (Ala, Arg, Cys, His, Pro, Ser, Tyr) have inside/outside preferences which are not consistent with their usual assignment as either polar or nonpolar residues; explanations are given for these apparent inconsistencies. Of the three types of backbone structure considered here (extended, alpha helix, and nonregular), extended structures have the greatest preference for the inside of proteins, and nonregular structures have the greatest preference for the outside. It is suggested that differences in entropy play an important part in the inside/outside preferences of backbone structures. There are generally significant changes in the conformational preferences of the residues in going from the inside to the outside of proteins; environmental (rather than local) solute-solvent interactions seem to be the predominant cause of these changes in conformational preferences.

21. Hydrophobicity parameters to predict bitter taste (Venanzi, 1984)
    H VENT840101
    D Bitterness (Venanzi, 1984)
    R LIT:1103107b PMID:6521488
    A Venanzi, T.J.
    T Hydrophobicity parameters and the bitter taste of L-amino acids
    J J. Theor. Biol. 111, 447-450 (1984)

    Abstract: The use of hydrophobicity parameters to predict the bitter taste of L-amino acids is discussed. It is concluded that no single hydrophobicity scale can be used to explain the complete range of L-amino acid behavior.

22. Slopes proteins FDPB VFF neutral (Avbelj, 2000)
    H AVBF000109
    D Slopes proteins, FDPB VFF neutral (Avbelj, 2000)
    R PMID:10903873

A Avbelj, F.
T Amino acid conformational preferences and solvation of polar backbone atoms
in peptides and proteins
J J. Mol. Biol. 300, 1335-1359 (2000) (Pro missing)

Abstract: Amino acids in peptides and proteins display distinct preferences for alpha-helical, beta-strand, and other conformational states. Various physicochemical reasons for these preferences have been suggested: conformational entropy, steric factors, hydrophobic effect, and backbone electrostatics; however, the issue remains controversial. It has been proposed recently that the side-chain-dependent solvent screening of the local and non-local backbone electrostatic interactions primarily determines the preferences not only for the alpha-helical but also for all other main-chain conformational states. Side-chains modulate the electrostatic screening of backbone interactions by excluding the solvent from the vicinity of main-chain polar atoms. The deficiency of this electrostatic screening model of amino acid preferences is that the relationships between the main-chain electrostatics and the amino acid preferences have been demonstrated for a limited set of six non-polar amino acid types in proteins only. Here, these relationships are determined for all amino acid types in tripeptides, dekapeptides, and proteins. The solvation free energies of polar backbone atoms are approximated by the electrostatic contributions calculated by the finite difference Poisson-Boltzmann and the Langevin dipoles methods. The results show that the average solvation free energy of main-chain polar atoms depends strongly on backbone conformation, shape of side-chains, and exposure to solvent. The equilibrium between the low-energy beta-strand conformation of an amino acid (anti-parallel alignment of backbone dipole moments) and the high-energy alpha conformation (parallel alignment of backbone dipole moments) is strongly influenced by the

solvation of backbone polar atoms. The free energy cost of reaching the alpha conformation is by approximately 1.5 kcal/mol smaller for residues with short side-chains than it is for the large beta-branched amino acid residues. This free energy difference is comparable to those obtained experimentally by mutation studies and is thus large enough to account for the distinct preferences of amino acid residues. The screening coefficients gamma(local)(r) and gamma(non-local)(r) correlate with the solvation effects for 19 amino acid types with the coefficients between 0.698 to 0.851, depending on the type of calculation and on the set of point atomic charges used. The screening coefficients gamma(local)(r) increase with the level of burial of amino acids in proteins, converging to 1.0 for the completely buried amino acid residues. The backbone solvation free energies of amino acid residues involved in strong hydrogen bonding (for example: in the middle of an alpha-helix) are small. The hydrogen bonded backbone is thus more hydrophobic than the peptide groups in random coil. The alpha-helix forming preference of alanine is attributed to the relatively small free energy cost of reaching the high-energy alpha-helix conformation. These results confirm that the side-chain-dependent solvent screening of the backbone electrostatic interactions is the dominant factor in determining amino acid conformational preferences. Copyright 2000 Academic Press.

23. Absolute entropy (Hutchens, 1970)

H HUTJ700102

D Absolute entropy (Hutchens, 1970)

R

A Hutchens, J.O.

T Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds

J In "Handbook of Biochemistry", 2nd ed. (Sober, H.A., ed.), Chemical Rubber

Co., Cleveland, Ohio, pp. B60-B61 (1970)

24. Retention coefficient in NaH2PO4 (Meek, 1981)
H MEEJ810102
D Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
R LIT:0708201
A Meek, J.L. and Rossetti, Z.L.
T Factors affecting retention and resolution of peptides in high-performance
liquid chromatography
J J. Chromatogr. 211, 15-28 (1981)

25. Linker index (Bae, 2005)
H BAEK050101
D Linker index (Bae et al., 2005)
R PMID:15746283
A Bae, K., Mallick, B.K. and Elsik, C.G.
T Prediction of protein inter-domain linker regions by a hidden Markov model
J Bioinformatics 21, ??-?? (2005)

Abstract: MOTIVATION: Our aim was to predict protein interdomain linker regions using sequence alone, without requiring known homology. Identifying linker regions will delineate domain boundaries, and can be used to computationally dissect proteins into domains prior to clustering them into families. We developed a hidden Markov model of linker/non-linker sequence regions using a linker index derived from amino acid propensity. We employed an efficient Bayesian estimation of the model using Markov Chain Monte Carlo, Gibbs sampling in particular, to simulate parameters from the posteriors. Our model recognizes sequence data to be continuous rather than categorical, and generates a probabilistic

output. RESULTS: We applied our method to a dataset of protein sequences in which domains and interdomain linkers had been delineated using the Pfam-A database. The prediction results are superior to a simpler method that also uses linker index.

26. Relative preference value at C' (Richardson, 1988)

H RICJ880116

D Relative preference value at C' (Richardson-Richardson, 1988)

R LIT:1408116 PMID:3381086

A Richardson, J.S. and Richardson, D.C.

T Amino acid preferences for specific locations at the ends of alpha helices

J Science 240, 1648-1652 (1988)

Abstract: A definition based on alpha-carbon positions and a sample of 215 alpha helices from 45 different globular protein structures were used to tabulate amino acid preferences for 16 individual positions relative to the helix ends. The interface residue, which is half in and half out of the helix, is called the N-cap or C-cap, whichever is appropriate. The results confirm earlier observations, such as asymmetrical charge distributions in the first and last helical turn, but several new, sharp preferences are found as well. The most striking of these are a 3.5:1 preference for Asn at the N-cap position, and a preference of 2.6:1 for Pro at N-cap + 1. The C-cap position is overwhelmingly dominated by Gly, which ends 34 percent of the helices. Hydrophobic residues peak at positions N-cap + 4 and C-cap - 4.

27. Normalized composition of mt proteins (Nakashima, 1990)

H NAKH900104

D Normalized composition of mt-proteins (Nakashima et al., 1990)

R LIT:2004138b PMID:2235995

A Nakashima, H., Nishikawa, K. and Ooi, T.
T Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
J Proteins 8, 173-178 (1990)

Abstract: A compact mitochondrial gene contains all essential information about the synthesis of mitochondrial proteins which play their roles in a small compartment of the mitochondrium. Almost no noncoding regions have been found through the gene, but a necessary set of tRNAs for the 20 amino acids is provided for biosynthesis, some of them coding different amino acids from those in a usual cell. Since the gene is so compact that the produced proteins would have some characteristic aspects for the mitochondrium, amino acid compositions of mitochondrial proteins (mt-proteins) were examined in the 20-dimensional composition space. The results show that compositions of proteins translated from the mitochondrial genes have a distinct character having more hydrophobic content than others, which is illustrated by a clustered distribution in the multidimensional composition space. The cluster is located at the tail edge of the global distribution pattern of a Gaussian shape for other various kinds of proteins in the space. The mt-proteins are rich in hydrophobic amino acids as is a membrane protein, but are different from other membrane proteins in a lesser content of Val. A good correlation found between the base and amino acid compositions for the mitochondria was examined in comparison to those of organisms such as thermophilic bacterium having an extreme G-C-rich base composition.

28. Linker propensity index (Suyama, 2003)
H SUYM030101
D Linker propensity index (Suyama-Ohara, 2003)
R PMID:12651735
A Suyama, M. and Ohara, O.

T DomCut: Prediction of inter-domain linker regions in amino acid sequences
J Bioinformatics 19, 673-674 (2003)

Abstract: DomCut is a program to predict inter-domain linker regions solely by amino acid sequence information. The prediction is made by using linker index deduced from a data set of domain/linker segments. The linker preference profile, which is the averaged linker index along a sequence, can be visualized in the graphical interface.

29. Consensus normalized hydrophobicity scale (Eisenberg, 1984)
H EISD840101
D Consensus normalized hydrophobicity scale (Eisenberg, 1984)
R LIT:2004004a PMID:6383201
A Eisenberg, D.
T Three-dimensional structure of membrane and surface proteins
J Ann. Rev. Biochem. 53, 595-623 (1984) Original references: Eisenberg, D., Weiss, R.M., Terwilliger, T.C. and Wilcox, W. Faraday Symp. Chem. Soc. 17, 109-120 (1982) Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity Proc. Natl. Acad. Sci. USA 81, 140-144 (1984)

30. Transfer free energy from chx to oct (Radzicka, 1988)
H RADA880104
D Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
R LIT:1405051b
A Radzicka, A. and Wolfenden, R.
T Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution
J Biochemistry 27, 1664-1670 (1988) (Pro Cys Asp missing)

31. Flexibility parameter for two rigid neighbors (Karplus, 1985)
H KARP850101
D Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)
R LIT:1110994
A Karplus, P.A. and Schulz, G.E.
T Prediction of chain flexibility in proteins
J Naturwiss. 72, 212-213 (1985)

32. Refractivity (McMeekin 1964, cited by Jones 1975)
H MCMT640101
D Refractivity (McMeekin et al., 1964), Cited by Jones (1975)
R
A McMeekin, T.L., Groves, M.L. and Hipp, N.J.
T
J In "Amino Acids and Serum Proteins" (Stekol, J.A., ed.), American Chemical
Society, Washington, D.C., p. 54 (1964)

33. Normalized composition from animal (Nakashima, 1990)
H NAKH900106
D Normalized composition from animal (Nakashima et al., 1990)
R LIT:2004138b PMID:2235995
A Nakashima, H., Nishikawa, K. and Ooi, T.
T Distinct character in hydrophobicity of amino acid composition of mitochondrial proteins
J Proteins 8, 173-178 (1990)

Abstract: A compact mitochondrial gene contains all essential information about the synthesis of mitochondrial proteins which play their roles in a small compartment of the mitochondrium. Almost no noncoding regions

have been found through the gene, but a necessary set of tRNAs for the 20 amino acids is provided for biosynthesis, some of them coding different amino acids from those in a usual cell. Since the gene is so compact that the produced proteins would have some characteristic aspects for the mitochondrium, amino acid compositions of mitochondrial proteins (mt-proteins) were examined in the 20-dimensional composition space. The results show that compositions of proteins translated from the mitochondrial genes have a distinct character having more hydrophobic content than others, which is illustrated by a clustered distribution in the multidimensional composition space. The cluster is located at the tail edge of the global distribution pattern of a Gaussian shape for other various kinds of proteins in the space. The mt-proteins are rich in hydrophobic amino acids as is a membrane protein, but are different from other membrane proteins in a lesser content of Val. A good correlation found between the base and amino acid compositions for the mitochondria was examined in comparison to those of organisms such as thermophilic bacterium having an extreme G-C-rich base composition.

34. Transfer free energy to surface (Bull, 1974)

H BULH740101

D Transfer free energy to surface (Bull-Breese, 1974)

R PMID:4839053

A Bull, H.B. and Breese, K.

T Surface tension of amino acid solutions: A hydrophobicity scale of the amino
acid residues

J Arch. Biochem. Biophys. 161, 665-670 (1974)

Abstract: The surface tensions of amino acids in 0.10 M NaCl have been measured as a function of the concentration of the amino acids at 30 °C using the method of Jones and Ray. From the experimental results, the

free energies of transfer of the amino acid residues from the solution to the surface have been calculated to yield a hydrophobicity scale of the residues. This scale is in fairly good agreement with that of Nozaki and Tanford. The temperature coefficient of the free energy of transfer of leucine residues to the surface gives about +10 entropy units per mole with an enthalpy of about +560 cal per mole at 30 °C.

35. Isoelectric point (Zimmerman, 1968)

H ZIMJ680104

D Isoelectric point (Zimmerman et al., 1968)

R LIT:2004109b PMID:5700434

A Zimmerman, J.M., Eliezer, N. and Simha, R.

T The characterization of amino acid sequences in proteins by statistical methods

J J. Theor. Biol. 21, 170-201 (1968)

Abstract: Three different but related comprehensive statistical analyses of amino acid sequences in proteins are described. The goal in each case is to search for evidence of significant sequence structure in individual proteins relative to a purely random arrangement of the amino acid residues and to attempt to relate any significant structure uncovered to the secondary and/or tertiary configuration of the protein.

In the first of these analyses, which is reviewed briefly in an appendix, amino acids are divided into subgroups according to a variety of side chain physical properties (e.g. polarity, hydrophobicity). Deviations from randomness are expressed in terms of correlation indices $\Omega_{ij}^{(c)}$ which are composition normalized doublet frequencies. Here *i* and *j* denote membership in a particular group for the physical property chosen and *c* denotes the “lag”, that is the number of residues along the chain separating the doublet.

The other more refined analyses are described in some detail. For both of these each amino acid in a given protein is replaced by its appropriate value on a continuous physical property scale. Six such scales are employed: bulkiness, polarity, R_F, p*I*, pK_1 and hydrophobicity. The resulting amino acid index sequences are treated as discrete series and are analyzed first by means of serial correlation methods and subsequently by employing spectral analysis techniques. Periodicities exhibited in these series are evaluated statistically and speculations are made concerning the connection between such structure and protein configuration.

Although more than forty individual proteins whose primary sequences are known have been analyzed by these methods, results for the cytochrome *c* series, the hemoglobins and lysozyme are emphasized in the present paper. In the case of the cytochrome *c* family of proteins several relationships between primary sequence structure and "evolutionary order" are discussed. In addition, the results of several homogeneity studies are described in which the sequence structure of various portions of a given protein chain are compared.

References

1. Wang, H. et al. Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *Journal of proteome research* **4**, 2397-2403 (2005).
2. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical chemistry* **75**, 663-670 (2003).
3. Ong, S.E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature protocols* **1**, 2650-2660 (2006).

4. Svetnik, V. et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences* **43**, 1947-1958 (2003).
5. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic acids research* **28**, 374 (2000).
6. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* **76**, 3908-3922 (2004).

Computational Method Comparison

In this study we compared the ESP predictor to existing methods designed to predict proteotypic peptides (Mallick *et al.*, Tang *et al.*, and Webb-Robertson *et al.*) Proteins were digested from two validation sets *in silico*, using trypsin, specifying a mass range of 600 – 2,800 Da and no missed cleavages. Depending on the program, either the peptide list or protein sequences were submitted to each program. When protein sequences were submitted only the peptides that fit the above criteria were considered, all others were ignored. The output for all programs is a probability. We ranked the peptides (high to low probability) for each protein based on this output and selected the top five. We then compared the predicted top five to the actual five highest responding peptides based on experimental data.

Mallick *et al.:* http://web.bii.a-star.edu.sg/~wongch/peptideSieve/
version 2008-May-29 (peptideSieve v0.51)

Tang *et al.*: http://proteomeartworks.com/

Webb-Robertson *et al.:* http://omics.pnl.gov/software/STEPP.php
Version 1.1; March 31, 2008