Supporting Information Figures and Tables

Figure S1. Significance model for energetic compatibility scoring. a) Gapless alignments between random amino acid sequences and strings of random thermodynamic environments follow length-dependent Gaussian distributions of total score. Randomly generated data are displayed as colored bars and analytical best-fit Gaussian distributions are shown as solid lines of the same color. Longer length alignments exhibit more negative average scores and larger variances. These distributions suggest that, regardless of length, energetically incompatible structures dominate the conformational space of a given amino acid sequence. b) Parameterization of a random model matching amino acid sequence to native state ensemble-based thermodynamic environments. Two simple mathematical models allow estimation of a random distribution of total score for an arbitrary length comparison. For a given length, the mean (μ) and standard deviation (σ) of the appropriate Gaussian can be accurately estimated, permitting estimation of the quality of the match (p-value). Gaussian PDF is defined as $f(x, \mu,\sigma) = (\sigma \, sqrt(2 \, \pi))^{-1} \, exp(-(x- \mu)^2 / 2 \, \sigma^2)$, where μ, σ are functions of length L.

Figure S2. Statistics of enrichment and depletion in extremely incompatible regions of sequence and structure. In all panels, error bars represent average and standard deviation of bootstrapping [37] the complete data into five equally-sized randomly selected subsets. "Highest Negative Compatibility Index" was defined as those regions of negative compatibility index greater than 75% of the complete data. a) Proline and Glycine are enriched in energetically incompatible regions of sequence, while polar / charged residues are depleted. b) Sequence gatekeeper positions exhibit similar trends as compared to panel a): Proline and Glycine are enriched in gatekeeper positions of sequence, while polar / charged residues are depleted. c) Stable structure positions, thermodynamic environments 6 – 8, are enriched in energetically incompatible regions of structure, while less stable regions are depleted.

d) Structure gatekeeper positions similarly show less depletion at the most stable positions and more depletion at the least stable positions, linking increased stability to increased gatekeeper propensity.

Table S1 (separate text file). Native state ensemble-based thermodynamic environments database of 122 human proteins of known structure used for this study. This database is identical to that previously published [44]. The index order corresponds to the rows and columns in Fig. 6 of the main text (for example, Protein 1, 1PBV, is located at lower left corner of Fig. 6). Columns denote Protein Data Bank (PDB) [33] identifier, *SCOP* database [32] classification, and number of residues in the PDB coordinate file.

Table S2 (separate text file). Four-dimensional densities of significantly compatible or incompatible energetic matches to all sequences and structures in the database. The first 10 columns of this Table are identical to that previously published [44], and are given to facilitate completeness and reproducibility of the analysis. Thermodynamic quantities (dG, dHap, dHpol, TdSconf) from the COREX / BEST algorithm [34] are given in values of kcal / mol under simulated native state conditions as described in Methods. "NCI" and "PCI" quantities listed in the next four columns correspond to values plotted in Fig. 7a&b, with the first 12 and last 13 residues omitted to ignore possible end-effects. "NCI" is the abbreviation for "Negative Compatibility Index" and "PCI" is the abbreviation for "Positive Compatibility Index". The final two columns, "SequenceClassification" and "Structure Classification" contain the classification of energetic compatibility described in Figure 3 of the text and plotted in Fig. 7c&d: "gatekeeper" as high NCI (greater than intra-protein median) and low PCI (less than or equal to intra-protein median), "permissive" as low PCI and high NCI, "selective" as high PCI and high NCI, and "inactive" as low NCI and low PCI.