tryptic peptide as well as the two flanking tryptic peptides to identify peptides with missed enzyme cleavage sites. Third, we applied the same procedure to single residue conflict annotations from Swiss-Prot (**Fig. 1**). This resulted in an increase in database size of about 10% (**Supplementary Discussion**). We also added proteolytic enzyme and keratin sequences. A second version of the database, MSIPI decoy, facilitates experimental analysis of significance scores[7]. MSIPI will be updated with each new IPI release, and is freely available for human and mouse from EBI (ftp://ftp.ebi.ac.uk/pub/databases/IPI/msipi). Tutorials on setup and customization as well as software scripts are available in **Supplementary Methods** and **Supplementary Software** online**. Figure 2** illustrates the process of constructing the database.

To test whether additional cSNPs, sequence conflicts and N-terminal peptides can be identified, we searched liquid chromatography–MS data from published and ongoing organellar proteomics projects against the MSIPI database generated from ipi.HUMAN.v3.23. Here we discuss the example of a proteomic investigation of lysosomes isolated from the human MCF7 breast cancer cell line and analyzed by two consecutive stages of mass spectrometry on a hybrid linear ion trap Fourier transform instrument (LTQ-FT) as described[8]. We found peptides for 13 cSNPs and 2 sequence conflicts in 13 proteins in the lysosomal data (see **Fig. 1c** for an analysis of one of the cSNP-containing peptides). This cSNP (valine to isoleucine change; refSNP identifier in ENSEMBL rs11549015) has frequencies ranging from 1% in Chinese to 21% in Caucasian populations. Sometimes we found both the peptide containing the most common allele and the peptide containing the polymorphism, indicating that both alleles are present in MCF7 cells (**Supplementary Table 1** online).

By searching the MSIPI database we also determined 13 N-terminal peptides from the lysosome proteome corresponding to peptides after removal of predicted signal peptides (**Supplementary Table 1**).

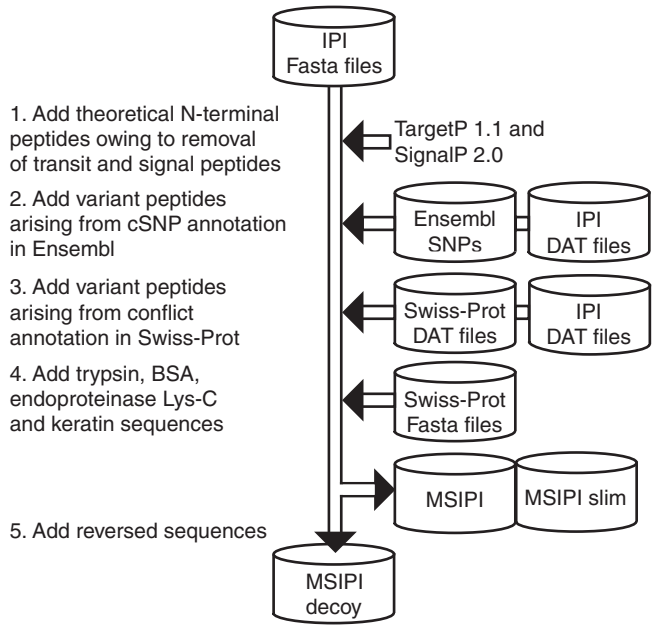

Figure 2 | Construction of the modified IPI database. The MSIPI database is generated in five steps which are described to the left and with the software and database dependencies listed to the right (see **Supplementary Methods**).

We identified a total of 29 additional peptides in the MSIPI database out of 7,832 peptides fulfilling stringent criteria, such as very small mass deviation (several parts per million), fully tryptic sequences, MS^2 score with at least 95% significance value and additional validation by a second (MS^3) fragmentation step[8]. We manually evaluated the peptide fragmentation spectra (**Supplementary Data** online).

In conclusion, the modified database described here allows efficient identification of N-terminal peptides and of cSNPs in proteomic samples, without substantially increasing the size of the database. Our scripts can easily be reconfigured for other databases. In the future, additional peptide variants could be included, such as frequently observed modifications and alternative splice sites not covered in IPI. We argue against wholesale incorporation of error-prone single-pass sequencing data from expressed sequence tag projects, which would substantially inflate the database and make validation difficult. With ongoing improvements in proteomic technology, detection of genetic polymorphisms on a large scale may become possible with MSIPI. This should provide a new dimension to the information gained in clinical proteomics projects.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank P. Kersey of the EMBL-European Bioinformatics Institute for distributing MSIPI from EBI. The Center for Experimental BioInformatics is supported by the Danish National Research Foundation and the Danish Platform for Integrative Biology.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Søren Schandorff[1], Jesper V Olsen[2], Jakob Bunkenborg[1], Blagoy Blagoev[1], Yong Zhang[2], Jens S Andersen[1] & Matthias Mann[1,2]

[1]Center for Experimental BioInformatics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. [2]Max Planck Institute of Biochemistry, Department of Proteomics and Signal Transduction, Am Klopferspitz 18, D-82152 Martinsried, Germany.
e-mail: mmann@biochem.mpg.de

1. Steen, H. & Mann, M. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
2. Kersey, P.J. *et al. Proteomics* **4**, 1985–1988 (2004).
3. Olsen, J.V., Ong, S.E. & Mann, M. *Mol. Cell. Proteomics* **3**, 608–614 (2004).
4. The International HapMap Consortium *Nature* **437**, 1299–1320 (2005).
5. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. *J. Mol. Biol.* **300**, 1005–1016 (2000).
6. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. *J. Mol. Biol.* **340**, 783–795 (2004).
7. Elias, J.E., Haas, W., Faherty, B.K. & Gygi, S.P. *Nat. Methods* **2**, 667–675 (2005).
8. Olsen, J.V. & Mann, M. *Proc. Natl. Acad. Sci. USA* **101**, 13417–13422 (2004).

Eris: an automated estimator of protein stability

To the editor: Mutagenesis is a central tool of molecular biology, genetics and biotechnology. To what extent mutations affect the thermodynamic stability ($\Delta\Delta G$) and structure of a protein is often vital for designing experiments. Estimation of protein stabilities remains a paramount challenge in computational molecular biology and has several bottlenecks. With recent advances in computational biology, accurate $\Delta\Delta G$ calculations can be achieved by molecular dynamic simulations[1], but such

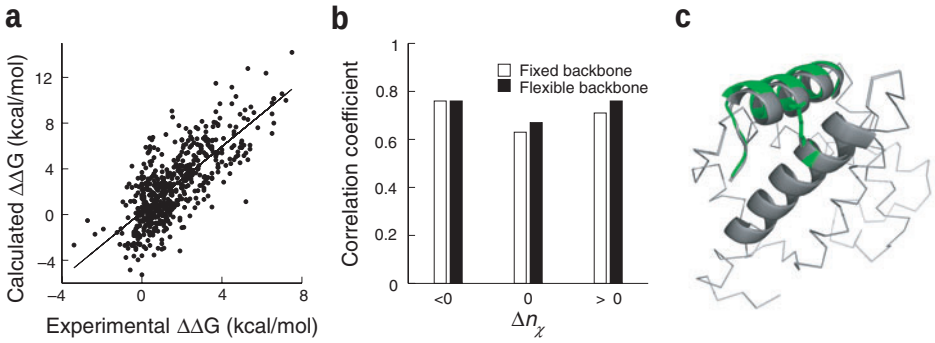

Figure 1 | Performance of Eris. (**a**) Scatter plot of $\Delta\Delta G$ calculations using Eris. The $\Delta\Delta G$ of 595 mutants were calculated and compared with experimental measurements. The Pearson correlation coefficient was 0.75 ($P = \sim 10^{-108}$) and the r.m.s. deviation between the experimental and computed $\Delta\Delta G$ values was 2.4 kcal/mol. The solid line corresponds to linear regression fit to the data points. (**b**) Correlation coefficients between the calculated and experimental $\Delta\Delta G$s for three different classes of mutations based on the change in the number of side-chain χ angles (Δn_χ). The mutations with $\Delta n_\chi < 0$ are associated with large-to-small mutations and those with $\Delta n_\chi \geq 0$ correspond to mutation to residues of the same or larger sizes. The flexible- and fixed-backbone methods have the same prediction accuracy for $\Delta n_\chi < 0$. However, the flexible-backbone $\Delta\Delta G$ prediction correlates better with experiments for $\Delta n_\chi \geq 0$ cases, owing to its ability to resolve possible side-chain clashes. (**c**) The backbone structures of wild-type and A130K mutant apomyoglobin proteins. The mutant structure is obtained from a flexible-backbone calculation. The N-terminal helix of the A130K apomyoglobin bends ~0.2 Å outward to accommodate the larger lysine side chain (green).

calculations are computationally costly (that is, efficiency is low). Modern large-scale $\Delta\Delta G$ prediction methods use heuristic algorithms with effective force fields and empirical parameters to estimate the stability changes caused by mutations in agreement with experimental data[2–5]. There are, however, two considerable drawbacks pertinent to the heuristic methods. First, most of these prediction methods rely on parameter training using available experimental $\Delta\Delta G$ data. Such training is usually biased toward mutations that feature large-to-small residue substitutions, such as alanine-scanning experiments (that is, poor transferability). Second, protein backbone flexibility, which is crucial for resolving atomic clashes and backbone strains in mutant proteins, is not considered in these methods, thereby reducing accuracy and limiting the application of heuristic methods (that is, limited applicability).

To address the issues of efficiency, transferability and applicability, we developed the Eris method, which uses a physical force field with atomic modeling as well as fast side-chain packing and backbone relaxation algorithms. The free energy is expressed as a weighted sum of van der Waals forces, solvation, hydrogen bonding and backbone-dependent statistical energies[6] (**Supplementary Methods** online). The weighting parameters are independently trained to recapitulate the native amino acid sequences for 34 proteins using high-resolution X-ray structures[6]. Additionally, an integral step of Eris is backbone relaxation when severe atom clashes or backbone strains are detected during calculation.

We tested Eris on 595 mutants from five proteins, for which the $\Delta\Delta G$ values were documented (**Fig. 1a**). We found significant agreement between the predicted and measured $\Delta\Delta G$ values with a correlation coefficient of 0.75 ($P = 2 \times 10^{-108}$). The correlation between the predictions and experiments is comparable to that reported using other methods[2–5]. Unlike previous methods, Eris also has high predictive power for small-to-large[3] side-chain-size mutations (**Fig. 1b**,**c**), owing to its ability to effectively relax backbone structures and resolve clashes introduced by mutations. As a direct comparison with other methods, we computed the stability changes of the small-to-large mutations using Eris and other web-based stability prediction servers. We found that Eris outperformed other available servers (**Supplementary Discussion** and **Supplementary Tables 1** and **2** online). Additionally, Eris features a protein structure pre-relaxation option, which remarkably improves the prediction accuracy when a high-resolution protein structure is not available (**Supplementary Discussion** and **Supplementary Fig. 1** online).

Our test validates the unbiased force field, side-chain packing and backbone relaxation algorithms in Eris. We anticipate Eris will be applicable to examining a much larger variety of mutations during protein engineering. We built a web-based Eris server for $\Delta\Delta G$ estimation. The server is freely accessible online (http://eris.dokhlab.org).

Note: Supplementary information is available on the Nature Methods website.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Shuangye Yin, Feng Ding & Nikolay V Dokholyan

Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
e-mail: dokh@med.unc.edu

1. Kollman, P. *Chem. Rev.* **93**, 2395–2417 (1993).
2. Gilis, D. & Rooman, M. *J. Mol. Biol.* **272**, 276–290 (1997).
3. Guerois, R., Nielsen, J.E. & Serrano, L. *J. Mol. Biol.* **320**, 369–387 (2002).
4. Bordner, A.J. & Abagyan, R.A. *Proteins* **57**, 400–413 (2004).
5. Saraboji, K. *et al. Biopolymers* **82**, 80–92 (2006).
6. Ding, F. & Dokholyan, N.V. *PLoS Comput. Biol.* **2**, e85 (2006).

An automated tool for maximum entropy reconstruction of biomolecular NMR spectra

To the editor: High resolution is essential for successful application of NMR spectroscopy to biomolecules, but involves a classic 'catch-22'. High magnetic fields increase chemical shift dispersion, thus increasing resolution and reducing spectral overlap, but the required increase in sampling rate (to avoid aliasing) means longer acquisition times in the indirect dimensions of multidimensional experiments (indirect dimensions are sampled by iteration, whereas the lone 'direct' dimension is sampled in real time). Consequently the potential resolution afforded by high magnetic fields is rarely realized in the indirect dimensions. There is a growing realization that this is a consequence of