Technologies for comments on an early draft of this publication.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Geoffrey P Lomax[1] & Alan O Trounson[1,2]

[1]California Institute for Regenerative Medicine, San Francisco, California, USA. [2]The Ritchie Centre, Monash Institute of Medical Research, Clayton, Australia.
e-mail: glomax@cirm.ca.gov

1. Goldstein, L. *J. Law Med. Ethics* **38**, 204–211 (2010).
2. Nordqvist, C. *Medical News Today* <http://www.medcalnewstoday.com/articles/247343.php> (2012).
3. CDC. 2010 National Summary, Assisted Reproductive Technology National Summary Report (Centers for Disease Control and Prevention, Atlanta, 2013).
4. Nachtigall, R.D. *et al. Fertil. Steril.* **92**, 2094–2096 (2009).
5. Bankowski, B.J., Lyerly, A.D., Faden, R.R. & Wallach, E.E. *Fertil. Steril.* **84**, 823–832 (2005).
6. Hoffman, D.I. *et al. Fertil. Steril.* **79**, 1063–1069 (2003).
7. Ertelt, S. LifeNews.com <http://www.lifenews.com/2012/03/05/obama-end-funding-for-snowflake-embryo-adoption-program/> (2012).
8. Palca, J. *Talk of the Nation* <http://www.npr.org/templates/story/story.php?storyId=96392644&ft=1&f=5> (2008).
9. Brezina, P.R. & Zhao, Y. *Obstet. Gynecol. Int.*, 686253 (2012).
10. Stein, J. *Milwaukee Journal Sentinel* (2012) <http://www.jsonline.com/news/statepolitics/how-thompson-baldwin-differ-on-stem-cell-research-im75rfj-174113021.html>.
11. Hug, K. *Fertil. Steril.* **89**, 263–277 (2008).
12. Hammarberg, K. & Tinney, L. *Fertil. Steril.* **86**, 86–91 (2006).
13. Lyerly, A.D. *et al. Fertil. Steril.* **93**, 499–509 (2008).
14. Kalista, T., Freeman, H.A., Behr, B., Pera, R.R. & Scott, C.T. *Cell Stem Cell* **8**, 360–362 (2011).
15. CIRM Medical and Ethical Standards Working Group July 25, 2008. <http://www.cirm.ca.gov/sites/default/files/files/agenda/transcripts/07-25-08.pdf>
16. International Stem Cell Registry. <http://www.umassmed.edu/iscr/index.aspx> (accessed 21 March 2013).
17. Patel, R. & Lomax, G.P. *J. Stem Cell Res. Ther.* **1**, 107 (2011).
18. Löser, P., Schirm, J., Guhr, A., Wobus, A.M. & Kurtz, A. *Stem Cells* **28**, 240–246 (2010).
19. Advisory, N.I.H. Committee to the Director GENEA Submission #2012-ACD-003. <http://acd.od.nih.gov/06142012_HeSC_002GENEA.pdf>.
20. Jacquet, L. *et al. EMBO Mol. Med.* **5**, 10–17 (2013).
21. Pruksananonda, K. *et al. BioResearch Open Access* **1**, 166–173 (2012).
22. CIRM New Cell Line Awards Request for Application, San Francisco, 2005. <http://www.cirm.ca.gov/our-funding/research-rfas/new-cell-lines> (accessed 21 March 2013).
23. CIRM. Medical and Ethical Standards Regulations. (California Institute for Regenerative Medicine, San Francisco, 2012).
24. National Research Council and Institute of Medicine. *Final Report of the National Academies' Human Embryonic Stem Cell Research Advisory Committee and 2010 Amendments to the National Academies' Guidelines for Human Embryonic Stem Cell Research* (National Academies Press, Washington, DC, 2010).

ProtoNet: charting the expanding universe of protein sequences

To the Editor:

As next-generation sequencing technologies continue to generate staggering amounts of raw protein sequences, it has become very difficult to thoroughly annotate the emerging protein-sequence space. Complete proteomes (that is, the collection of all valid proteins from a sequenced genome) as well as partial sequencing efforts have resulted in the archiving of more than 20 million protein sequences in UniProtKB (release 2012_1, 25 January 2012; http://www.uniprot.org). This repository is compiled from millions of viral sequences, thousands of microbial genomes and sequences from thousands of multicellular organisms. These sequences comprise what may be considered the now-known parts of the protein space. At present, the functional characterization available for the vast majority of this space is based mostly on sequence-similarity approaches. In fact, the characterized part of this space is orders of magnitude smaller than the whole, and only 3.5% of sequences in UniProtKB[1] have any experimental support. From this view, only a robust, unsupervised and automated method can realistically achieve comprehensive and functional annotation of this rapidly expanding protein space.

Protein three-dimensional structures provide the most reliable information on biochemical function. At present, there are 80,000 solved protein structures (http://www.rcsb.org/pdb/home/home.do) that are indirectly associated with a large fraction of the protein space. Through semi-automatic classifications, these three-dimensional solved proteins are organized in an inventory of ~1,500 basic folds[2,3]. However, these folds are consistent with local domains rather than full-length proteins. Complementary sequence–based approaches for protein family assignment rely primarily on the notion of domains as the building blocks of proteins. The general scheme starts with multiple sequence alignment, which is then translated into statistically based models (e.g., Pfam)[4]. The integration of different resources (e.g., InterPro)[5] leads to a substantial increase in domain coverage of the protein space. The curated portion of the protein space is already enormous. Still, one-third (6.7 million) of UniProtKB sequences are marked as putative or hypothetical. For these sequences, current methods for direct inference of function with high confidence have mostly failed. Furthermore, most sequence- and structure-based assignments rely on local information such as structural fold, sequence domain and functional signature). Consequently, functional annotations at the level of the full-length protein are prone to erroneous inference. It is realistic to expect an even faster growth in the number of protein sequences (e.g., from large-scale sequencing of environmental samples). This creates a pressing need for accurate methods of annotation inference.

We offer the ProtoNet 6.1 family tree (http://www.protonet.cs.huji.ac.il), a classification resource created by an unsupervised analysis of protein sequences[6]. The families in the ProtoNet tree are generated through the following steps: (i) precalculation of sequence-similarity values for all possible pairwise relationships (all against all BLAST values), (ii) application of an unsupervised bottom-up clustering algorithm (this algorithm organizes large sets of proteins in a hierarchical tree that yields high-quality protein families) (**Supplementary Table 1**) and (iii) a process of pruning the ProtoNet tree to retain only the most informative clusters. This computational process yields a tree-like skeleton of the entire known protein space. In the next stage, each cluster is assessed through a comprehensive battery of descriptors for domains, three-dimensional structures, enzymes, gene ontology, taxonomy and more (**Supplementary Tables 2** and **3**). In addition, rigorous annotation-based quality tests are carried out to assign a statistically based quality measure for each stable cluster. Each cluster is then assigned the set of descriptors that reflect the most significant annotation(s) of its proteins. In this way, ProtoNet circumvents many of the pitfalls in annotation inference discussed above.

There are several features that allow ProtoNet to cope with the scale of the known protein space. First, it applies a scalable, efficient and accurate algorithm for clustering millions of sequences[7]. Second, family construction is 'model free'; the process of tree construction is continuous and data driven. Third, all sequences are dealt with on an equal basis, irrespective of length, domain organization, taxonomy or prior knowledge. Thus, putative proteins play an integral part in the construction of ProtoNet.

Nearly 19 million full-length protein sequences are included in ProtoNet 6.1

(**Supplementary Text**). Thus, the most up-to-date ProtoNet database provides an almost 7.8-fold expansion of the representatives as defined by UniRef50 (ref. 1). **Figure 1** summarizes the main features and principles involved in constructing the ProtoNet tree. The algorithm is an unsupervised, bottom-up agglomerative averaging protocol that outperforms naive algorithms used toward this task (**Fig. 1a**). A pruned tree retains the most robust and reliable clusters representing the evolutionary relatedness of protein sequences. From the biological, functional-inference approach, ProtoNet captures, at varying levels of granularity, the functional relations between subfamilies and superfamilies (**Fig. 1b**). Most importantly, distances in the graph metric of in the tree provide an intrinsic approximation of evolutionary relatedness. Therefore, the tree exposes the often-overlooked evolutionary relatedness of isolated families. ProtoNet uses all leading annotation resources (for structure, sequence, function and taxonomy) for inference (**Supplementary Table 2**). Thus, each cluster is assigned a name that is derived from the most informative annotation(s) for its proteins (ProtoName; **Fig. 1c** and **Supplementary Fig. 2**). At a resolution that ensures high-quality annotation, there are 150,000 robust clusters that nevertheless cover the entire protein-sequence space (**Fig. 1d** and **Supplementary Fig. 1**). Overall, ProtoNet infers function for a huge portion of the undercharacterized 'putative' and 'hypothetical' proteins at a wide range of cluster sizes (**Fig. 1e**,**f**). Hence, ProtoNet 6.1 reduces the chasm between the full-length protein sequences that have been acquired and the still-uncharted complexity of their evolutionary origins.

ProtoNet is useful to the biomedical community for knowledge extraction and for navigation toward new discoveries. We emphasize the gain in knowledge for clusters that best capture annotations from various resources. The correspondence score (CS; **Supplementary Text**) quantifies both the purity and the coverage of each selected annotation. **Figure 2** shows the fraction of functional inference according to the major annotation resources that are covered by ProtoNet. **Figure 2a** shows an analysis of 2,069 specific annotations that matched the highest-CS clusters, partitioned according to the source of annotation. On average, 15.6% of the proteins in these clusters were newly annotated by ProtoNet (**Supplementary Tables 5** and **6**). One such cluster (**Fig. 2b**) includes proteins from more than 100 bacteria, including pathogens *Haemophilus influenzae*, *Neisseria meningitidis* (a meningitis-causing bacterium) and *Yersinia enterocolitica* (which causes a zoonotic disease occurring in humans, cattle and birds). Navigation of ProtoNet from a sequence from *Y. enterocolitica* (Q70W78, putative uncharacterized protein open reading frame 7) shows the sequence of mergers that yield a cluster (size 98; **Fig. 2b**) of maximal quality for the InterPro term 'addiction module antidote protein, HI1420' (CS = 1.0). Additional mergers did not reduce the quality of the root cluster (121 representative proteins; **Fig. 2b**). Only one protein in this cluster is reviewed (UniProtKB/Swiss-Prot quality tag; **Supplementary Fig. 3**). In the expanded cluster (371 proteins; **Supplementary Table 6**) most sequences are labeled 'putative', 'possible', 'predicted', 'hypothetical' or 'uncharacterized' (**Fig. 2c**). Still, the ProtoName considers the InterPro term shared by the most proteins in the cluster (53 out of 121 proteins) as the most probable automatic inference for the proteins in the cluster. All proteins that carry this term are included in the root cluster with no false negatives. This test case is indicative for thousands of instances that originate from newly sequenced or unannotated genomes (**Supplementary Table 6**).

Aside from inferences regarding particular sequences, ProtoNet navigation tools allow us to trace high-quality clusters even when the biological information associated with the clusters' sequences is mostly missing. About 25% of safe inferences concerning clusters are annotated as domains of unknown function (DUF)[4] (**Fig. 2d** and **Supplementary Table 9**). We can thus focus on mergers in which clusters grow in size without losing quality.

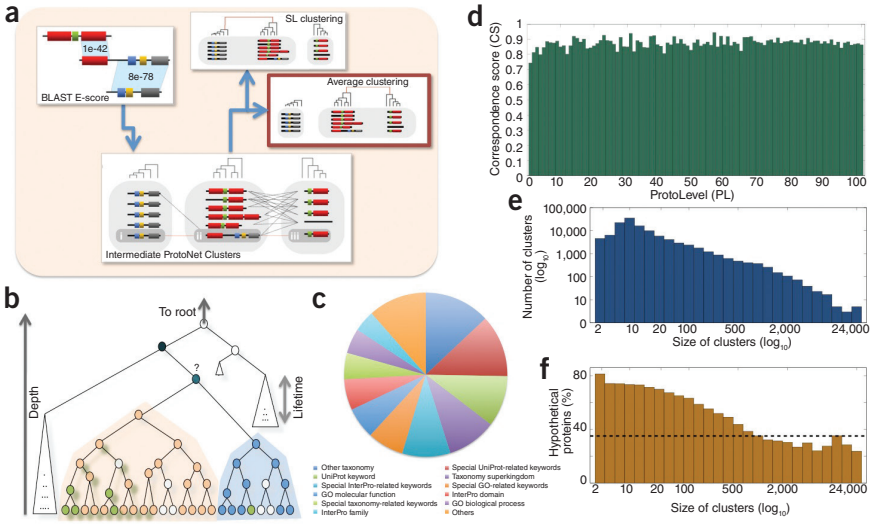

Figure 1 The ProtoNet family tree. (**a**) The clustering protocol is applied using pre-calculated all-against-all BLAST E-scores (from 0 to 100). The CM (constrained memory) ProtoNet algorithm[7] generates a binary tree with 2.5 million nodes (UniProt50 representatives). Average clustering outperforms the single-linkage algorithm (**Fig. 1a–c**). (**b**) New connections are proposed by the merging of high-quality clusters (demonstrated by the two subtrees shaded light orange and light blue). Open circles denote proteins of putative/hypothetical proteins. Functional inference is suggested from the annotation of the clustered proteins. A scheme allows the assessment of potentially overlooked connections between clusters (indicated by a question mark connecting the two high-quality subtrees). (**c**) The contribution of annotation types to ProtoNet clusters. ProtoNet uses 40 different annotation types. Major annotation types are listed. In ProtoNet 6.0, 144 million annotations (74.4 million excluding taxonomy) are associated with ~9 million protein sequences. (**d**) Quality assessment of Pfam keyword assignments. The cluster with the best average CS for each Pfam keyword (for clusters of more than 20 proteins) is shown for ProtoLevels 0–100. The quality of the clusters (according to CS) is very high at all PLs. (**e**) Number of clusters by cluster size at ProtoLevel (PL = 0.95, $\log_{10}$ scale). At this PL, there are 120,000 clusters ranging in size from 2 to 24,000 proteins per cluster. (**f**) The percentage of putative hypothetical proteins in clusters of sizes indicated. The fraction of hypothetical proteins in small-sized clusters can reach 80%. At least 33% of the proteins in clusters of 2,000–20,000 proteins are putative hypothetical proteins. The dashed line indicates the percentage of putative hypothetical proteins in the entire database.

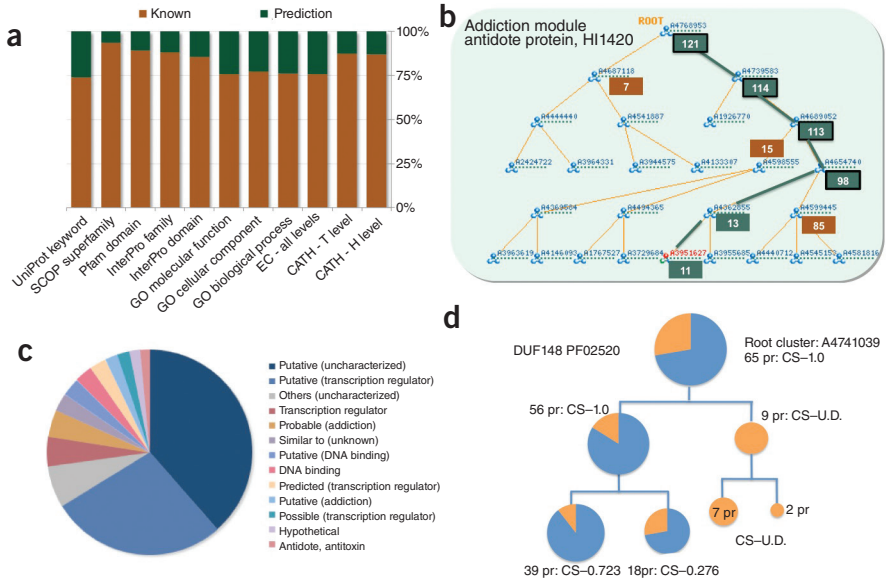

Figure 2 Navigating the ProtoNet tree. (**a**) The knowledge gain for each annotation type is indicated by the percentage of known (that is, already annotated) and prediction (that is, ProtoNet inference) for all clusters with the maximal CS values. A maximal CS cluster is a cluster that receives the highest CS value for a specific annotation. The knowledge gain for all annotation types is 15.6% (based on 2,069 annotations covering ~144,000 representative proteins (**Supplementary Table 7**). (**b**) Snapshot of the ProtoNet tree for protein Q70W78 from *Y. enterocolitica*. Nodes represent clusters. The sequence was merged with a stable cluster of 11 proteins that lack functional annotation (ProtoLevel (PL) = 37.6, LT = 37). LifeTime (LT) of a cluster is the difference between PL at its creation (that is, the time when two clusters were merged to form the present cluster) and its termination (that is, the time where the cluster was merged with another cluster). The LT of a cluster reflects its remoteness from the clusters in its 'vicinity'. Therefore, LT is an intrinsic measure that approximates the stability of a cluster. The merge that created a cluster of 98 proteins (PL = 91.8) has a CS = 1.0 and was marked as best cluster for InterPro term 'addiction module antidote protein, HI1420' (53 annotated proteins). Numbers in boxes indicate the number of proteins in each cluster. Clusters along the branch of Q70W78 are shown in green. The main clusters merged with that branch are shown in orange. Framed rectangles denotes clusters with CS = 1.0 for addiction module antidote protein, HI1420. Only stable clusters (LT > 1.0) are shown. (**c**) Root cluster A4768953, containing 121 representatives from a wide range of bacteria, expanded to 371 proteins (**Supplementary Tables 5–8** and **Supplementary Fig. 3**). Most proteins in this cluster are labeled 'putative', 'similar', 'predicted', 'hypothetical' or 'probable'. (**d**) The subtree of the root cluster A4741039 is an example of a DUF cluster that has a maximal CS value. Cluster along the merging process are indicated (circles). Blue indicates the fraction of proteins annotated Pfam DUF148; orange indicates the fraction of proteins that lack annotations. The CS of the clusters is indicated. U.D., undefined CS; pr, protein. Among the 65 proteins, half are marked 'putative' and a large number are marked 'identified transcript' (proteins are listed in **Supplementary Table 9**).

Several design principles were implemented to cope with the fast growth in the number of protein sequences and even faster accumulation of associated annotations (**Supplementary Table 4**). Comparison of the quality of ProtoNet 6.1 to that of previous versions (**Supplementary Table 1**) reveals that the quality of the clusters remains very high in the face of increasing volume (**Supplementary Fig. 1**).

ProtoNet is scalable and can cope with the fast growth in the protein space. We increased the coverage of the database from 9 million sequences (in version 6.0) to 18.9 million sequences (in version 6.1, based on UniProtKB release 2012_1, 25 January 2012). Most (6 million) of the sequences not yet covered are bacterial proteins.

ProtoNet constitutes an intuitive navigational tool with which to explore the entire protein space. Its highlighting of overlooked connections between families, division of families into subfamilies and reliable inference constitute immediate benefits to the biological and biomedical community.

Note: Supplementary information is available at http://dx.doi.org/10.1038/nbt.2553.

ACKNOWLEDGMENTS

We thank S. Karsenty for support in design and maintenance of the ProtoNet server. We thank A. Stern for extensive analysis on the ProtoNet performance. This work was supported by an European Commission Seventh Framework Programme Prospects grant.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests in the online version of the paper (doi:10.1038/nbt.2553).

Nadav Rappoport[1], Nathan Linial[1] & Michal Linial[2]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. [2]Department of Biological Chemistry, Institute of Life Sciences, The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Jerusalem, Israel.
e-mail: michall@cc.huji.ac.il

1. The UniProt Consortium. *Nucleic Acids Res.* **40**, D71–D75 (2012).
2. Cuff, A.L. *et al. Nucleic Acids Res.* **37**, D310–D314 (2009).
3. Andreeva, A. *et al. Nucleic Acids Res.* **32**, D226–D229 (2004).
4. Finn, R.D. *et al. Nucleic Acids Res.* **36**, D281–D288 (2008).
5. Hunter, S. *et al. Nucleic Acids Res.* **37**, D211–D215 (2009).
6. Rappoport, N. *et al. Nucleic Acids Res.* **40**, D313–D320 (2012).
7. Loewenstein, Y. *et al. Bioinformatics* **24**, i41–i49 (2008).

Alternative splicing and protein interaction data sets

To the Editor:

Alternative splicing, which produces several isoforms of the same protein from a single gene, is extremely common in higher eukaryotes. Splicing events often lead to enormous differences among isoforms in their sequences and structures and in the interactions formed. The differences in protein-protein interactions (PPIs) between different isoforms are generally overlooked when data are made publicly available and can lead to both false-positive and false-negative interactions in large-scale data sets. Thus, we advocate here that the