Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 398968, 6 pages
http://dx.doi.org/10.1155/2013/398968

Research Article

Cloud Prediction of Protein Structure and Function with PredictProtein for Debian

László Kaján,[1] Guy Yachdav,[1,2,3] Esmeralda Vicedo,[1] Martin Steinegger,[1] Milot Mirdita,[1] Christof Angermüller,[1] Ariane Böhm,[1] Simon Domke,[1] Julia Ertl,[1] Christian Mertes,[1] Eva Reisinger,[1] Cedric Staniewski,[1] and Burkhard Rost[1,2,3,4,5]

[1] *TUM, Department of Informatics, Bioinformatics & Computational Biology-I12, Boltzmannstraß 3, 85748 Garching, Germany*
[2] *Columbia University, Department of Biochemistry and Molecular Biophysics and New York Consortium on Membrane Protein Structure (NYCOMPS), 701 West 168th Street, New York, NY 10032, USA*
[3] *Biosof LLC, 10th Floor, 138 West 25th Street, New York, NY 10001, USA*
[4] *WZW-Weihenstephan, Alte Akademie 8, Freising, Germany*
[5] *Institute for Advanced Study (TUM-IAS), Lichtenbergstraß 2a, 85748 Garching, Germany*

Correspondence should be addressed to László Kaján; lkajan@rostlab.org

Received 4 January 2013; Accepted 5 July 2013

Academic Editor: Ching-Hsien (Robert) Hsu

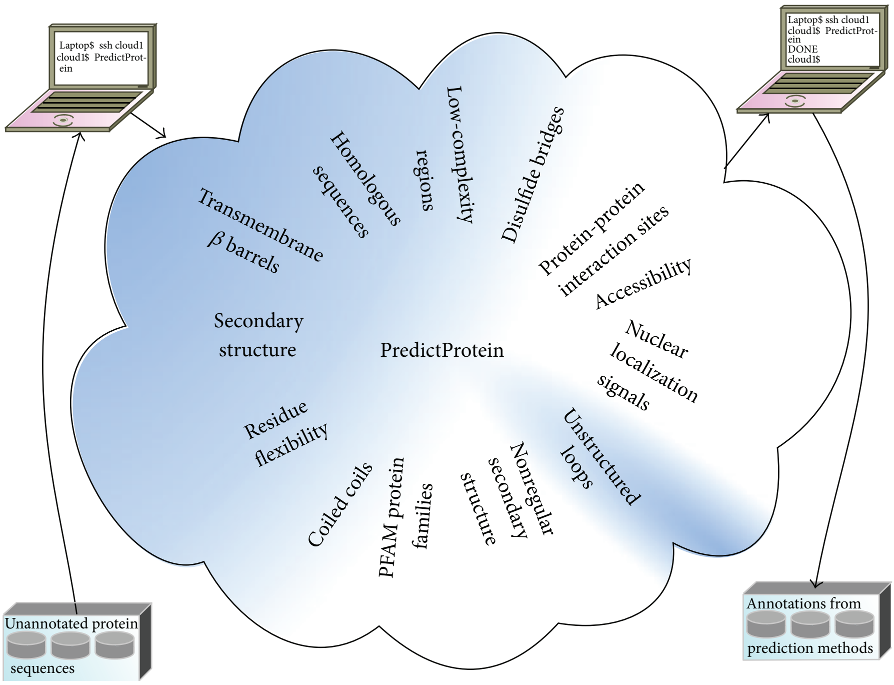
We report the release of PredictProtein for the Debian operating system and derivatives, such as Ubuntu, Bio-Linux, and Cloud BioLinux. The PredictProtein suite is available as a standard set of open source Debian packages. The release covers the most popular prediction methods from the Rost Lab, including methods for the prediction of secondary structure and solvent accessibility (profphd), nuclear localization signals (predictnls), and intrinsically disordered regions (norsnet). We also present two case studies that successfully utilize PredictProtein packages for high performance computing in the cloud: the first analyzes protein disorder for whole organisms, and the second analyzes the effect of all possible single sequence variants in protein coding regions of the human genome.

1. Background

Bioinformatics is embracing cloud computing. Recent months have seen the publication of cloud sequence analysis platforms, CloVR [1] and Galaxy Cloud [2], and the cloud version of Bio-Linux [3], Cloud BioLinux [4]. Cost analysis depicts cloud computing as an attractive and sustainable solution for computational biology and bioinformatics [5–8]. The rate of data generation of "next generation" sequencing (NGS) drives the efforts to turn to cloud computing as a solution to handling peak-time loads, without the need to maintain large clusters [9]. Cloud-enabled bioinformatics tools are now available in the context of high throughput sequencing and genomics [10].

The Rost Lab provides protein structure and function prediction tools for cloud computing in the PredictProtein suite [11]. PredictProtein began as an Internet server for sequence analysis and the prediction of aspects of protein structure and function in 1992 [12]. Queried with a protein sequence, PredictProtein returns secondary structure and accessibility predictions, predictions of unstructured loops, nuclear localization signals, protein-protein interaction sites, disulfide bonds, regions lacking regular secondary structure, protein family hits, low-complexity regions, bacterial transmembrane beta barrels, coiled-coil regions, protein residue flexibility, and homologous sequences (Figure 1).

Cloud computing is commonly realized on machine instances that run on virtual hardware providing "infrastructure as a service" (IaaS) [13, 14]. This type of cloud computing instantiates compute nodes from machine images. Machine images usually contain an operating system with software tools. For example, one could request the instantiation of 10 worker nodes of PredictProtein on Debian operating system at the Amazon EC2 IaaS offering.



```
Laptop$  ssh cloud1
cloud1$  PredictProt-
ein
```

```
Laptop$ ssh cloud1
cloud1$  PredictProt-
ein
DONE
cloud1$
```

Figure 1: Protein annotation by PredictProtein. PredictProtein annotates input sequences with the features shown.

The PredictProtein cloud solution builds upon the open source operating system Debian [15] and provides its functionality as a set of free [16] software packages. Bio-Linux is an operating system for bioinformatics and computational biology. The latest Bio-Linux release 7 provides more than 500 bioinformatics programs on an Ubuntu Linux base [17]. Ubuntu is a "derivative" operating system [18] based on Debian, with its own additions. Cloud BioLinux is a comprehensive cloud solution that is derived from Bio-Linux and Ubuntu. Debian derivatives can easily share packages between each other. For example, Debian packages are automatically incorporated in Ubuntu [19] and are also usable in Cloud BioLinux (the procedure is described in [4]).

2. Implementation

The PredictProtein suite is implemented as a set of free packages released at http://debian.org/. Software packaging conformed with the Policy Manual [20], and following the recommendations of the Developer's Reference [21].

3. Results and Discussion

High-throughput experiments generate vast amounts of data at an ever-increasing rate; the pace of creating reliable annotations needed to use that data increases much slower. One of the major challenges for computational tools is to narrow the resulting increase in the protein annotation gap [22]. Of the over 35 m (million) sequences in the UniProt Knowledgebase 2013_05 [23], only about 500 k (500 thousand) have explicit experimental annotations in Swiss-Prot [24]. Computational prediction methods, such as those included in PredictProtein, can annotate important features for the remainder and enable us to draw scientific insights. Unfortunately, the task is often intractable for any single desktop computer within reasonable time. Fortunately, cloud computing is now at hand. On-demand servers in the cloud promise to fit computing power to most tasks economically, and without a fair portion of the usual worries of system management: hardware purchasing, recruiting a system manager, high availability issues, and so forth ([13] and the references therein). One problem remains: how to get the often adhoc analysis toolset from the desktop environment into the cloud? Directly addressing this problem, here we report the first Debian package release of the protein feature prediction toolset "PredictProtein," developed at the Rost Lab.

The publication of scientific results has, overall, changed surprisingly little since the Internet exists [25]. Research code is regularly distributed as a "zip" file of the development directory. Often, the only "documentation" distributed along with the code is the published paper accompanied by some "README" file. Software distributed this way often fails outside the laboratory without expert attention. In order to address this issue in the PredictProtein suite, we decided to apply the community and time-tested packaging and release requirements of Debian to PredictProtein components. We have traced all dependencies, eliminated convenience copies, carefully documented each of our prediction methods, and made them go through the thorough review process every Debian package receives. This converted PredictProtein from an adhoc implementation to a reusable software component (Figure 2).

Our packages facilitate the generation of purpose-built machine images for cloud computing. As an example, we distribute a slim PredictProtein machine image (PPMI) through

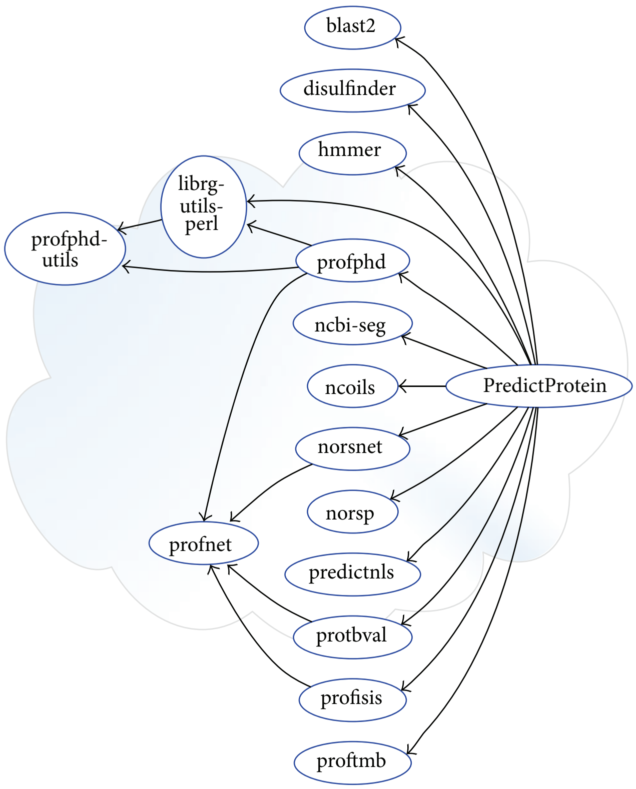


Figure 2: Package dependencies for PredictProtein. Arrows represent "depends on" relationships. Only significant dependencies are shown for clarity. Convenience copies of "profnet" for "profphd," "norsnet," "profbval," and "profisis" have been merged to a single "profnet" package. Similar merging was done for all code convenience copies.

the PredictProtein website [26]. This image contains a minimal installation of Debian with the command line version of PredictProtein. Databases are provided as a separate disk image. The PPMI is bootable on server instances in cloud infrastructure services, or on locally installed virtualization software. The latter allows for a cross-platform solution to use PredictProtein. Apart from virtualization, "chroot" environments present an option to run the software on Linux distributions where Debian packages are not readily usable. After booting the machine image, a friendly message at the login prompt offers usage tips and directions to documentation. A "Getting Started with PredictProtein" guide is available online [27]. The PPMI and the data image are updated regularly and are freely available at http://predictprotein.org/. For a comprehensive bioinformatics and computational biology computing environment, we recommend using PredictProtein with Bio-Linux [3] or Cloud BioLinux [4], where PredictProtein is either preinstalled or is easily installable from package repositories. We plan to release the web-based graphical interface of PredictProtein for these platforms in the near future.

The PredictProtein suite has attracted respectable popularity both online and offline. PredictProtein has been operating continuously since 1992, that is, the dawn of the Internet. Today, over 100,000 online users are registered; over 500 users access the PredictProtein web page every day and 12,000 unique users apply the service every month. Our Media Wiki page presenting an overview of the Rost Lab software packages has been accessed nearly 60,000 times since its launch 36 months ago. Adoption of the PredictProtein packages by the community has also been remarkable. Over 200 packages of the PredictProtein suite are installed from the Debian repository alone, while these and other installations have performed over 57 million protein feature predictions over the past year, not counting our own usage. Out of this, ~30 million were secondary structure and accessibility predictions from the "profphd" method [28].

4. PredictProtein Packages

The following protein feature prediction methods—components of PredictProtein—are available (feature—"package name"): secondary structure, accessibility, and transmembrane helices—"profphd" [29–31]; unstructured loops—"norsnet" [32]; nuclear localization signals—"predictnls" [33]; protein-protein interaction sites—"profisis" [34]; disulfide bridges—"disulfinder" [35]; nonregular secondary structure—"norsp" [36]; PFAM hits—"hmmer" [37, 38]; local complexity—"ncbi-seg" [39]; bacterial transmembrane beta barrels—"proftmb" [40]; coiled-coils—"ncoils" [41]; protein residue flexibility—"profbval" [42]; sequence homologies—"blast2" [43]; protein feature prediction suite—"predictprotein" [11].

These tools are available under a free license through Debian and are automatically incorporated into other Linux distributions such as Ubuntu. An overview of the packages offered for bioinformatics and cloud computing, complete with literature references, is available at Debian Med [44]. PredictProtein is listed in the Biology task.

5. Case Study 1: Protein Disorder in Completely Sequenced Organisms

The goal of this study is to collect evidence for three hypotheses on protein disorder: (1) it is more useful to picture disorder as a distinct phenomenon than as an extreme example of protein flexibility; (2) there are many very different flavors of protein disorder, but it is advantageous to recognize just two main types, namely, *well structured* and *disordered*; (3) nature uses protein disorder as a tool to adapt to different environments [45]. We predicted protein disorder both on an inhouse compute grid and on a compute grid manually setup in the OpenNebula [46] cloud service provided by the CSC Finland [47]. Data and tool (the PPMI) images for grid nodes in the cloud were downloaded from http://predictprotein.org/. The PPMI image was extended with a grid client, and a separate machine instance was used as grid master. PredictProtein for the local grid was installed from the main Debian repository. Required databases (28 GB) were included on a data disk image for cloud machine instances. Input to PredictProtein jobs consisted of protein sequences (in total less than 1 GB). Grid job submissions to the local and the cloud grid were manually adjusted according to available resources. Over 9 million disorder predictions were made over the course of the past few years.

6. Case Study 2: Comprehensive In Silico Mutagenesis of Human Proteome

This project aims at providing information about the functional effect of every possible point mutation in all human proteins, that is, for the replacement of $19*N$ amino acids for a protein with N residues. Overall, this generated 300 million human sequence variants (point mutants). The method SNAP [48] predicted the effect of each variant, that is, each "nonsynonymous single nucleotide polymorphisms" (nsSNPs) upon protein function. These predictions are useful for several reasons. First, the study of all possible mutations in human will provide the background against which we can assess the effect of mutations that are actually observed between people. This is crucial for both the advance toward personalized medicine and health and the understanding of human diversity and variation. Second, our computation provides quick "lookup" answers available for all the important variants that are observed and implied in important phenotypes. The only way to cover those lookups is by precomputing all the possible changes. SNAP can take advantage of PredictProtein results for faster processing. With the PredictProtein packages presented here, a solution was built in the form of a public Amazon Machine Image (AMI, ami-3f5f8156) that allows running PredictProtein on the Amazon Elastic Compute Cloud (EC2). We extended an Ubuntu-based StarCluster [49] AMI with PredictProtein and its required databases (28 GB). Because every protein can be computed independently, we formed a grid job out of each protein and used the Grid Engine (GE) to distribute work on the machine instances. We used StarCluster to automate grid setup on the EC2. Because a lot of CPU power was needed, the "Cluster Compute Eight Extra Large Instance" was chosen. This instance type is especially crafted for big data with a lot of CPU power. One instance has 60.5 GB memory, 88 EC2 Compute Units (2x Intel Xeon E5-2670, eight-core-architecture "Sandy Bridge"), and 3370 GB instance storage. The sequence variants were analyzed based on the human reference proteome from the National Center for Biotechnology Information (build 37.3, proteins, 21MB). We processed 29,036 sequences with 16,618,608 residues. This amounted to predicting the functional effect of 315,753,552 individual amino acid changes.

7. Conclusion

The open source release of the PredictProtein protein structure and function prediction suite from the Rost Lab is now available for Debian and derivative operating systems, such as Ubuntu, Bio-Linux, and Cloud BioLinux. The software, due to its standard packaging, is readily deployable in the cloud. Successfully addressing the challenges of cloud computing brings PredictProtein—developed over almost two decades—into the present and the future. In accordance with the Rost Lab open policy [50], and supported by anonymous statistics, PredictProtein is now shared with a wide range of users. We encourage the bioinformatics community to take advantage of our open source software, itself a result of the collaboration of the wider open source software community.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contributions

L. Kaján and G. Yachdav (equal contributors) have redesigned "predictprotein," performed initial software packaging, and wrote the paper; E. Vicedo, M. Steinegger and M. Mirdita performed case studies, and reviewed the paper; C. Angermüller, A. Böhm, S. Domke, J. Ertl, C. Mertes, E. Reisinger, and C. Staniewski finalized the packaging for Debian; B. Rost provided initial implementation of the "predictprotein" core module and reviewed the paper.

Acknowledgments

This work was supported by the Alexander von Humboldt Foundation, the IAS at the TU in Munich and the National Institutes of Health (R01-LM07329 to G. Yachdav). M. Steinegger and M. Mirdita received funding from Amazon and Cycle Computing for case study 2. The authors thank all those who funded our research. The authors thank researchers who deposited data into publicly available datasets and programmers who provided their work under a free license: our work stands upon their shoulders and would not have been possible without them. The authors wish to thank the Debian project in general and Steffen Möller and Andreas Tille in particular for their tireless support.

References

[1] S. V. Angiuoli, M. Matalka, A. Gussman et al., "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC Bioinformatics*, vol. 12, p. 356, 2011.

[2] E. Afgan, D. Baker, N. Coraor et al., "Harnessing cloud computing with Galaxy Cloud," *Nature Biotechnology*, vol. 29, no. 11, pp. 972–974, 2011.

[3] D. Field, B. Tiwari, T. Booth et al., "Open software for biologists: from famine to feast," *Nature Biotechnology*, vol. 24, no. 7, pp. 801–803, 2006.

[4] K. Krampis, T. Booth, B. Chapman et al., "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics*, vol. 13, p. 42, 2012.

[5] J. T. Dudley, Y. Pouliot, R. Chen, A. A. Morgan, and A. J. Butte, "Translational bioinformatics in the cloud: an affordable alternative," *Genome Medicine*, vol. 2, no. 8, p. 51, 2010.

[6] S. V. Angiuoli, J. R. White, M. Matalka, O. White, and W. F. Fricke, "Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing," *PLoS ONE*, vol. 6, no. 10, Article ID e26624, 2011.

[7] P. Kudtarkar, T. F. DeLuca, V. A. Fusaro, P. J. Tonellato, and D. P. Wall, "Cost-effective cloud computing: a case study using the comparative genomics tool, roundup," *Evolutionary Bioinformatics*, vol. 2010, no. 6, pp. 197–203, 2010.

[8] M. Steinegger, *HPC Full in Silico Mutagenesis, in Department of Bioinformatics*, Technical University of Munich, Munich, Germany, 2012.

[9] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, p. 207, 2010.

[10] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy CloudMan: delivering cloud compute clusters," *BMC Bioinformatics*, vol. 11, supplement 12, p. S4, 2010.

[11] B. Rost and J. Liu, "The PredictProtein server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3300–3304, 2003.

[12] B. Rost, G. Yachdav, and J. Liu, "The PredictProtein server," *Nucleic Acids Research*, vol. 32, pp. W321–W326, 2004.

[13] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, pp. 50–55, 2008.

[14] V. A. Fusaro, P. Patil, E. Gafni, D. P. Wall, and P. J. Tonellato, "Biomedical cloud computing with amazon web services," *PLoS Computational Biology*, vol. 7, no. 8, Article ID e1002147, 2011.

[15] J. J. Amor, G. Robles, J. M. González-Barahona, and I. Herraiz, "From pigs to stripes: a travel through debian," in *Proceedings of the Debian Annual Developers Meeting (DebConf '05)*, Citeseer, 2005.

[16] "The Debian Free Software Guidelines (DFSG)," http://www.debian.org/social_contract#guidelines.

[17] B. T. Dawn Field, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston, "Bio-Linux 7," http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-7-info, 2012.

[18] "Debian Derivatives," http://wiki.debian.org/Derivatives.

[19] "NEW packages through Debian," https://wiki.ubuntu.com/UbuntuDevelopment/NewPackages#NEW_packages_through_Debian.

[20] "Debian Policy Manual," http://www.debian.org/doc/debian-policy/, 2012.

[21] "Debian Developer's Reference," http://www.debian.org/doc/manuals/developers-reference/, 2012.

[22] Y. Bromberg, G. Yachdav, Y. Ofran, R. Schneider, and B. Rost, "New in protein structure and function annotation: hotspots, single nucleotide polymorphisms and the "Deep Web"," *Current Opinion in Drug Discovery and Development*, vol. 12, no. 3, pp. 408–419, 2009.

[23] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, p. bar009, 2011.

[24] A. Bairoch, B. Boeckmann, S. Ferro, and E. Gasteiger, "Swiss-Prot: juggling between evolution and stability," *Briefings in Bioinformatics*, vol. 5, no. 1, pp. 39–55, 2004.

[25] R. Gentleman, "Reproducible research: a bioinformatics case study," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, p. 1034, 2005.

[26] "PredictProtein Website," http://predictprotein.org/.

[27] L. Kajan, "Getting Started with PredictProtein," http://wiki.debian.org/DebianMed/PredictProtein, 2013.

[28] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 16, pp. 7558–7562, 1993.

[29] B. Rost and C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," *Proteins*, vol. 19, no. 1, pp. 55–72, 1994.

[30] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, no. 3, pp. 216–226, 1994.

[31] B. Rost, R. Casadio, P. Fariselli, and C. Sander, "Transmembrane helices predicted at 95% accuracy," *Protein Science*, vol. 4, no. 3, pp. 521–533, 1995.

[32] A. Schlessinger, J. Liu, and B. Rost, "Natively unstructured loops differ from other loops," *PLoS Computational Biology*, vol. 3, no. 7, p. e140, 2007.

[33] M. Cokol, R. Nair, and B. Rost, "Finding nuclear localization signals," *EMBO Reports*, vol. 1, no. 5, pp. 411–415, 2000.

[34] Y. Ofran and B. Rost, "ISIS: interaction sites identified from sequence," *Bioinformatics*, vol. 23, no. 2, pp. e13–e16, 2007.

[35] A. Ceroni, A. Passerini, A. Vullo, and P. Frasconi, "Disulfind: a disulfide bonding state and cysteine connectivity prediction server," *Nucleic Acids Research*, vol. 34, pp. W177–W181, 2006.

[36] J. Liu and B. Rost, "NORSp: predictions of long regions without regular secondary structure," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3833–3835, 2003.

[37] R. D. Finn, J. Mistry, J. Tate et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D211–D222, 2010.

[38] S. R. Eddy, "Accelerated profile HMM searches," *PLoS Computational Biology*, vol. 7, no. 10, Article ID e1002195, 2011.

[39] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers and Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[40] H. Bigelow and B. Rost, "PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins," *Nucleic Acids Research*, vol. 34, pp. W186–W188, 2006.

[41] A. Lupas, "[30] Prediction and analysis of coiled-coil structures," *Methods in Enzymology*, vol. 266, pp. 513–524, 1996.

[42] A. Schlessinger, G. Yachdav, and B. Rost, "PROFbval: predict flexible and rigid residues in proteins," *Bioinformatics*, vol. 22, no. 7, pp. 891–893, 2006.

[43] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[44] S. Möller, H. N. Krabbenhöft, A. Tille et al., "Community-driven computational biology with Debian Linux," *BMC Bioinformatics*, vol. 11, supplement 12, p. S5, 2010.

[45] A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta, and B. Rost, "Protein disorder–a breakthrough invention of evolution?" *Current Opinion in Structural Biology*, vol. 21, no. 3, pp. 412–418, 2011.

[46] R. Moreno-Vozmediano, R. Montero, and I. Llorente, "IaaS cloud architecture: from virtualized data centers to federated cloud infrastructures," *IEEE Computer Society*, vol. 45, no. 12, pp. 65–72, 2012.

[47] T. H. Nyrönen, M. A. Babar, C. E. Cuesta, and J. E. Savolainen, "Delivering ICT infrastructure for biomedical research," in *Proceedings of the WICSA/ECSA 2012 Companion Volume*, pp. 37–44, ACM, Helsinki, Finland, 2012.

[48] Y. Bromberg and B. Rost, "SNAP: predict effect of non-synonymous polymorphisms on function," *Nucleic Acids Research*, vol. 35, no. 11, pp. 3823–3835, 2007.

[49] "StarCluster," http://star.mit.edu/cluster/index.html.

[50] R. H. Lathrop and B. Rost, "ISCB public policy statement on open access to scientific and technical research literature," *Bioinformatics*, vol. 27, no. 3, pp. 291–294, 2011.