Editorial:
Building successful biological databases

In this issue, we are happy to present a collection of four articles that discuss the key factors in building a successful biological database. The bioinformatics community does many things, but we can roughly summarise most activities as either building algorithms or building databases. Successful algorithms are measured in terms of their time and space complexity, and there are well-known standards for communicating and validating new algorithms. Databases are much more difficult to validate. How do we know that a database is good? Do we judge it by the quality or quantity of its data, its usability and its ability to integrate with outside resources? Should we also consider the technical architecture and the user support mechanisms and documentation? What about the impact on science, the number of web hits, and total gigabytes of data that a database transfers? Of course, we probably should include all these factors in deciding if a database is a success.

Anyone who builds and maintains biological databases also knows that it is very hard to publish traditional academic papers about databases. Familiar refrains include 'What’s the hypothesis?', 'What’s the evaluation scheme?', and ‘This is not science, this is engineering’. In partial response to this difficulty, there is the annual special ‘database issue’ of *Nucleic Acids Research*, and the occasional Application Note in the back of the *Bioinformatics* journal, and other similar publications. But these articles are always severely limited in page length, and can sometimes read as database advertisements, written to attract potential users.

We believe that *Briefings in Bioinformatics* has an opportunity to provide a forum for the maintainers of successful database projects to publish analyses of their approach, and to reveal the general lessons they have learned about the task of creating a successful biological database. Although many issues in creating a good database may transcend biology and be valid for all domains, there are special circumstances around biological databases that make them worth treating as a special group (eg the presumptions that they be freely available, that they be on the internet, that they keep up with a rapidly growing field and that they maintain high biological relevance).

To assemble this special issue, we solicited manuscripts from the scientists associated with a shortlist of databases that the editors and their consultants felt could be called ‘successful’ with little controversy. These ranged over a variety of data types, but concentrated on databases dealing with molecular biology, genomics and functional genomics. We encouraged authors to write papers about the technical engineering of their databases, and the other factors that they thought were significant. We asked them to address issues such as getting ‘buy-in’ from their user communities, designing user interfaces, creating software middle-layers for data management, integrating data from multiple sources, gathering and responding to user feedback, and setting policies for data curation and provenance.

We are thrilled that four of the database teams we contacted agreed to contribute to this issue: the *Saccharomyces* Genome Database (SGD), the *Ensembl* suite of databases, the Protein Data Bank (PDB) and the *Swiss-Prot* protein sequence database. We could hardly have asked for a more representative sampling of databases, in terms of their
scientific missions and histories. In the four reports included in this issue, the database teams cover a broad range of important challenges. We will let you read and discover the striking commonalities across the experiences building these databases, and the interesting differences. However, we cannot help but observe that every one of these reports stresses the non-technical aspects of creating and maintaining a successful database. To be sure, there are important challenges and lessons on the technical software side and the biological domain side, but these challenges seem to pale in comparison with the formidable sociological challenges associated with convincing scientists to share their data at a level of detail rarely matched in the published literature.

We hope that this special issue contributes to a critically important literature on the social and technical engineering of good biological databases. This literature will be useful to all of us in the database business, and may help define a set of ‘best practices’ that can be used to build the next generation of databases. In the last decade, biologists have shown an impressive ability to create high-throughput data collection methods, and have also demanded flexible and powerful ways to organise, store, integrate and analyse multiple data types. The four reports presented together here constitute our current best understanding of how to meet this challenge. A brave bioinformatics professional embarking on the creation of a new database for biomedical science can read the reports in this special issue, and ponder the hard-won lessons they present.

Russ B. Altman
Stanford University
Guest Editor, Special Issue