บทความที่น่าสนใจประจำเดือนมิถุนายน 2557
สาขาวิทยาศาสตร์และเทคโนโลยี

1	Title:	Wavefunction methods for the accurate characterization of water clusters
	Author:	Howard, J. C. and Tschumper, G. S.
	Journal:	Wiley Interdisciplinary Reviews: Computational Molecular Science, Volume 4, Issue 3, pages 199–224, May/June 2014
	Abstract:	Although the first ab initio Hartree–Fock computations of the water dimer were reported more than four decades ago, the detailed characterization of water clusters with sophisticated electronic structure techniques remains an important and vibrant area of research. The field of computational quantum chemistry has made significant advances since those pioneering studies. Geometry optimizations of the water dimer can now be carried out at the CCSDTQ level, and CCSD(T) energies can be computed with the aug-cc-pVTZ basis for clusters as large as (H2O)17. Some of these high-level studies are starting to reveal that the electronic structure is harder to describe for some hydrogen bonds than others. For example, discrepancies between MP2 and CCSD(T) energetics tend to increase when there are qualitative differences in the hydrogen-bonding networks of the water clusters being studied. This review highlights the recent and exciting work in this area and provides an overview of popular strategies for generating reliable properties and benchmark quality energetics for water clusters with correlated wavefunction methods.
	Database:	Wiley Online Library

2	Title:	Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles
	Author:	Riera, C., Lois, S. and de la Cruz, X.
	Journal:	Wiley Interdisciplinary Reviews: Computational Molecular Science, Volume 4, Issue 3, pages 249–268, May/June 2014
	Abstract:	The recent drop in genome sequencing costs has created a promising horizon for the development of genomic medicine. Within the biomedical environment, sequencing data are increasingly used for disease diagnosis and prognosis, treatment development, counseling, and so on. Many of these applications rely on the identification of disease causing variants. This is a particularly challenging problem because of the large number and wide variety of sequence variants identified in sequencing projects, and also because we only have a limited understanding of the physicochemical/biochemical properties that differentiate neutral from pathologic variants. Nonetheless, these last years have witnessed important methodological advances for one class of variants, those corresponding to changes in the amino-acid sequence of proteins. Proteins are a main constituent of living systems. We

	know that although their biological properties are essentially determined by the amino-acid sequence, not all the changes in this sequence have the same impact. Some are neutral, but others affect protein function and lead to disease. A large body of evidence shows that whether one or the other is the case that depends on properties such as mutation location in the protein structure, interspecies conservation, and so on. Mutation prediction methods based on these features have good success rates, in the 70–90% range, although representation over time suggests there is a performance plateau that would limit their applicability. In light of the most recent advances in the field, and after reviewing the foundations of prediction methods, we discuss the existence of this performance threshold and how it can be overcomed.
Database:	Wiley Online Library

3	Title:	Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation
	Author:	Baker, C. M. and Best, R. B.
	Journal:	Wiley Interdisciplinary Reviews: Computational Molecular Science, Volume 4, Issue 3, pages 182–198, May/June 2014
	Abstract:	Intrinsically disordered proteins (IDPs) are a class of protein that, in the native state, possess no well-defined secondary or tertiary structure, existing instead as dynamic ensembles of conformations. They are biologically important, with approximately 20% of all eukaryotic proteins disordered, and found at the heart of many biochemical networks. To fulfil their biological roles, many IDPs need to bind to proteins and/or nucleic acids. And although unstructured in solution, IDPs typically fold into a well-defined three-dimensional structure upon interaction with a binding partner. The flexibility and structural diversity inherent to IDPs makes this coupled folding and binding difficult to study at atomic resolution by experiment alone, and computer simulation currently offers perhaps the best opportunity to understand this process. But simulation of coupled folding and binding is itself extremely challenging; these molecules are large and highly flexible, and their binding partners, such as DNA or cyclins, are also often large. Therefore, their study requires either simplified representations, advanced enhanced sampling schemes, or both. It is not always clear that existing simulation techniques, optimized for studying folded proteins, are well suited to IDPs. In this article, we examine the progress that has been made in the study of coupled folding and binding using molecular dynamics simulation. We summarize what has been learnt, and examine the state of the art in terms of both methodologies and models. We also consider the lessons to be learnt from advances in other areas of simulation and highlight the issues that remain of be addressed.
	Database:	Wiley Online Library

4	Title:	Comparison of approaches for parameter identifiability analysis of biological systems
	Author:	Andreas Raue, Johan Karlsson, Maria Pia Saccomani, Mats Jirstrand, and Jens Timmer
	Journal:	Bioinformatics, Volume 30 Issue 10 May 15, 2014, Pp. 1440-1448
	Abstract:	Motivation: Modeling of dynamical systems using ordinary differential equations is a popular approach in the field of Systems Biology. The amount of experimental data that are used to build and calibrate these models is often limited. In this setting, the model parameters may not be uniquely determinable. Structural or a priori identifiability is a property of the system equations that indicates whether, in principle, the unknown model parameters can be determined from the available data. Results: We performed a case study using three current approaches for structural identifiability analysis for an application from cell biology. The approaches are conceptually different and are developed independently. The results of the three approaches are in agreement. We discuss strength and weaknesses of each of them and illustrate how they can be applied to real world problems. Availability and implementation: For application of the approaches to further applications, code representations (DAISY, Mathematica and MATLAB) for benchmark model and data are provided on the authors webpage.
	Database:	Oxford Journals Online

5	Title:	Modeling tissue contamination to improve molecular identification of the primary tumor site of metastases
	Author:	Martin Vincent, Katharina Perell, Finn Cilius Nielsen, Gedske Daugaard, and Niels Richard Hansen
	Journal:	Bioinformatics, Volume 30 Issue 10 May 15, 2014, Pp. 1417-1423
	Abstract:	Motivation: Contamination of a cancer tissue by the surrounding benign (non-cancerous) tissue is a concern for molecular cancer diagnostics. This is because an observed molecular signature will be distorted by the surrounding benign tissue, possibly leading to an incorrect diagnosis. One example is molecular identification of the primary tumor site of metastases because biopsies of metastases typically contain a significant amount of benign tissue. Results: A model of tissue contamination is presented. This contamination model works independently of the training of a molecular predictor, and it can be combined with any predictor model. The usability of the model is illustrated on primary tumor site identification of liver biopsies, specifically, on a human dataset consisting of microRNA expression measurements of primary tumor samples, benign liver samples and liver metastases. For a predictor trained on primary tumor and benign liver samples, the contamination model decreased the test error on biopsies from liver metastases from 77 to 45%. A further reduction to 34% was obtained by including biopsies in the training data.
	Database:	Oxford Journals Online

6

Title:	G-BLASTN: accelerating nucleotide alignment by graphics processors
Author:	Kaiyong Zhao and Xiaowen Chu
Journal:	Bioinformatics, Volume 30 Issue 10 May 15, 2014, Pp. 1384-1391
Abstract:	Motivation: Since 1990, the basic local alignment search tool (BLAST) has become one of the most popular and fundamental bioinformatics tools for sequence similarity searching, receiving extensive attention from the research community. The two pioneering papers on BLAST have received over 96 000 citations. Given the huge population of BLAST users and the increasing size of sequence databases, an urgent topic of study is how to improve the speed. Recently, graphics processing units (GPUs) have been widely used as low-cost, high-performance computing platforms. The existing GPU-BLAST is a promising software tool that uses a GPU to accelerate protein sequence alignment. Unfortunately, there is still no GPU-accelerated software tool for BLAST-based nucleotide sequence alignment. Results: We developed G-BLASTN, a GPU-accelerated nucleotide alignment tool based on the widely used NCBI-BLAST. G-BLASTN can produce exactly the same results as NCBI-BLAST, and it has very similar user commands. Compared with the sequential NCBI-BLAST, G-BLASTN can achieve an overall speedup of 14.80X under 'megablast' mode. More impressively, it achieves an overall speedup of 7.15X over the multithreaded NCBI-BLAST running on 4 CPU cores. When running under 'blastn' mode, the overall speedups are 4.32X (against 1-core) and 1.56X (against 4-core). G-BLASTN also supports a pipeline mode that further improves the overall performance by up to 44% when handling a batch of queries as a whole. Currently G-BLASTN is best optimized for databases with long sequences. We plan to optimize its performance on short database sequences in our future work.
Database:	Oxford Journals Online

7

Title:	Alignment-free phylogenetics and population genetics
Author:	Bernhard Haubold
Journal:	Briefings in Bioinformatics, Volume 15 Issue 3 May, 2014, Pp. 407-418
Abstract:	Phylogenetics and population genetics are central disciplines in evolutionary biology. Both are based on comparative data, today usually DNA sequences. These have become so plentiful that alignment-free sequence comparison is of growing importance in the race between scientists and sequencing machines. In phylogenetics, efficient distance computation is the major contribution of alignment-free methods. A distance measure should reflect the number of substitutions per site, which underlies classical alignment-based phylogeny reconstruction. Alignment-free distance measures are either based on word counts or on match lengths, and I apply examples of both approaches to simulated and real data to assess their accuracy and efficiency. While phylogeny reconstruction is based on the number of substitutions, in population genetics, the distribution of mutations along a sequence is also

	considered. This distribution can be explored by match lengths, thus opening the prospect of alignment-free population genomics.
Database:	Oxford Journals Online

8	Title:	Knowledge representation in metabolic pathway databases
	Author:	Miranda D. Stobbe, Gerbert A. Jansen, Perry D. Moerland, and Antoine H.C. van Kampen
	Journal:	Briefings in Bioinformatics, Volume 15 Issue 3 May, 2014, Pp. 455-470
	Abstract:	The accurate representation of all aspects of a metabolic network in a structured format, such that it can be used for a wide variety of computational analyses, is a challenge faced by a growing number of researchers. Analysis of five major metabolic pathway databases reveals that each database has made widely different choices to address this challenge, including how to deal with knowledge that is uncertain or missing. In concise overviews, we show how concepts such as compartments, enzymatic complexes and the direction of reactions are represented in each database. Importantly, also concepts which a database does not represent are described. Which aspects of the metabolic network need to be available in a structured format and to what detail differs per application. For example, for in silico phenotype prediction, a detailed representation of gene–protein–reaction relations and the compartmentalization of the network is essential. Our analysis also shows that current databases are still limited in capturing all details of the biology of the metabolic network, further illustrated with a detailed analysis of three metabolic processes. Finally, we conclude that the conceptual differences between the databases, which make knowledge exchange and integration a challenge, have not been resolved, so far, by the exchange formats in which knowledge representation is standardized.
	Database:	Oxford Journals Online

9	Title:	HeteroGenome: database of genome periodicity
	Author:	Maria Chaley, Vladimir Kutyrkin, Gayane Tulbasheva, Elena Teplukhina, and Nafisa Nazipova
	Journal:	Database: The Journal of Biological Databases and Curation, Volume 2014, doi: 10.1093/database/bau040
	Abstract:	We present the first release of the HeteroGenome database collecting latent periodicity regions in genomes. Tandem repeats and highly divergent tandem repeats along with the regions of a new type of periodicity, known as profile periodicity, have been collected for the genomes of Saccharomyces cerevisiae, Arabidopsis thaliana, Caenorhabditis elegans and Drosophila melanogaster. We obtained data with the aid of a spectral-statistical approach to search for reliable latent periodicity regions (with periods up to 2000 bp) in DNA sequences. The original two-level mode of data presentation (a broad view of the region of latent periodicity and a second level indicating conservative fragments of its structure) was further developed to enable us to obtain the estimate, without redundancy, that latent

	periodicity regions make up 10% of the analyzed genomes. Analysis of the quantitative and qualitative content of located periodicity regions on all chromosomes of the analyzed organisms revealed dominant characteristic types of periodicity in the genomes. The pattern of density distribution of latent periodicity regions on chromosome unambiguously characterizes each chromosome in genome.
Database:	Oxford Journals Online

10

Title:	The eGenVar data management system—cataloguing and sharing sensitive data and metadata for the life sciences
Author:	Sabry Razick, Rok MoČnik, Laurent F. Thomas, Einar Ryeng, Finn Drabløs, and Pål Sætrom
Journal:	Database: The Journal of Biological Databases and Curation, Volume 2014, doi: 10.1093/database/bau027
Abstract:	Systematic data management and controlled data sharing aim at increasing reproducibility, reducing redundancy in work, and providing a way to efficiently locate complementing or contradicting information. One method of achieving this is collecting data in a central repository or in a location that is part of a federated system and providing interfaces to the data. However, certain data, such as data from biobanks or clinical studies, may, for legal and privacy reasons, often not be stored in public repositories. Instead, we describe a metadata cataloguing system and a software suite for reporting the presence of data from the life sciences domain. The system stores three types of metadata: file information, file provenance and data lineage, and content descriptions. Our software suite includes both graphical and command line interfaces that allow users to report and tag files with these different metadata types. Importantly, the files remain in their original locations with their existing access-control mechanisms in place, while our system provides descriptions of their contents and relationships. Our system and software suite thereby provide a common framework for cataloguing and sharing both public and private data.
Database:	Oxford Journals Online