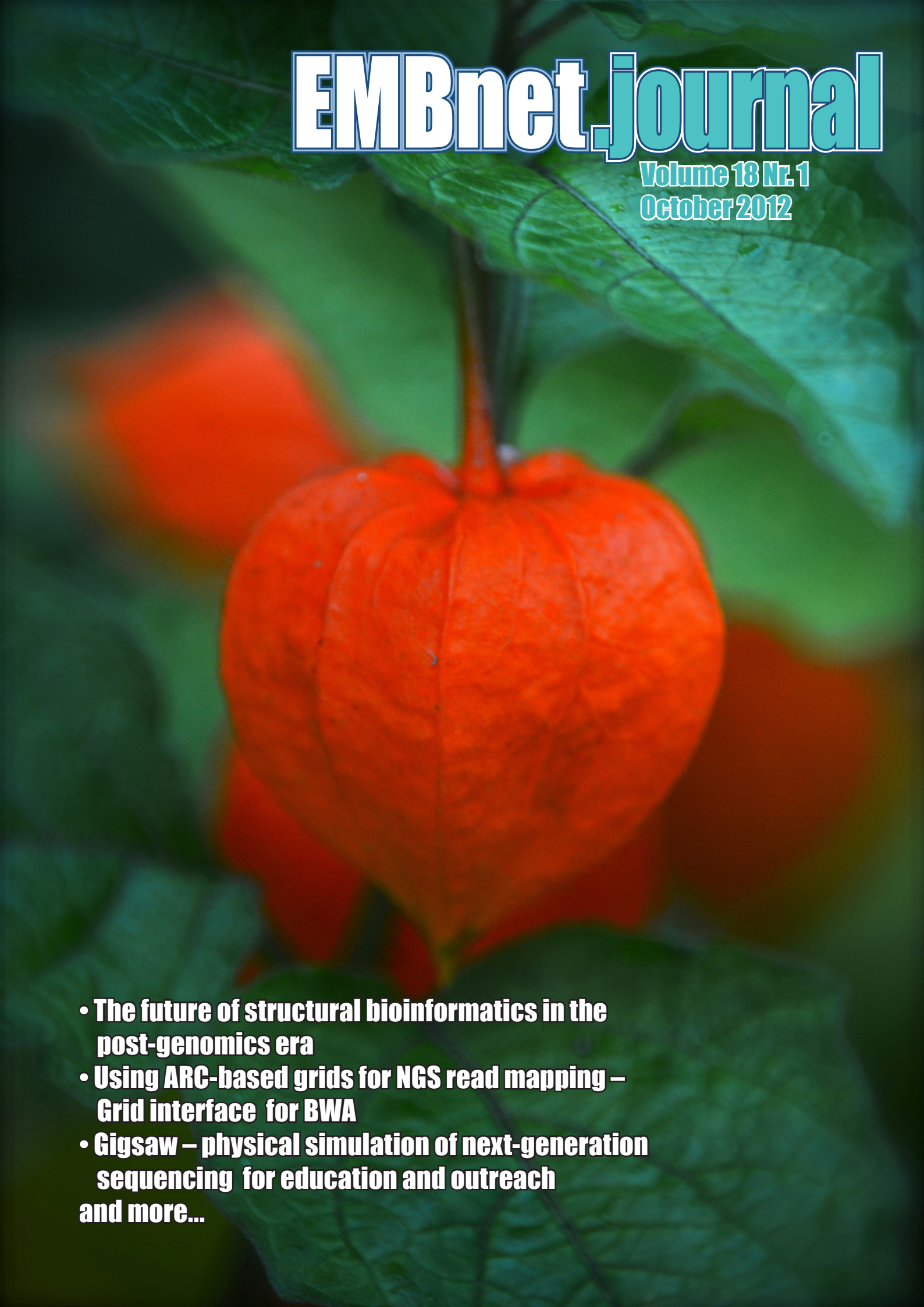


EMBnet.journal

Volume 18 Nr. 1
October 2012

- The future of structural bioinformatics in the post-genomics era
- Using ARC-based grids for NGS read mapping – Grid interface for BWA
- Gigsaw – physical simulation of next-generation sequencing for education and outreach

and more...

Editorial

The other day, I visited my usual hair-dresser to get a hair cut. She wasn't there that day, and so, having no other free time to attend on another day, I decided to ask one of the other hair-dressers in the salon.

As this was a new member of staff, she initiated the usual chat that we all have to endure during our regular hair-cut sessions: the weather, local events, and then the inevitable, "What do you do for a living?" I told her I work in the field of Bioinformatics. "Whats that?" she asked. I explained, in the most pedagogical way I could, and ended with, "We try to hunt down gene differences that make us what we are - tall or short, dark or blonde, sick or healthy." She looked at me with a happy face and replied, "One of my friends used to be a hair-dresser, but now she has a company that does exactly that!"

Surprised by the comment, I asked, "Is she in Bioinformatics?" "I don't know," she replied, "but she's doing tests that show whether people are carrying genes that make them fat or not. I'm planning to take one because, if a have a fat gene, why torture myself with diets if they don't work?!"

This little true-life story shows how fast things are moving in the Life Sciences, and that many Bioinformatics tools are nowadays used in the most incredible ways in the most unlikely places.

Our journal is playing, every day, a more important role in educating people far beyond the traditional research communities by embracing the Open Access ideology, which has the power to open a whole treasure-trove of knowledge to mankind. Our journal is also encouraging the submission of Educational articles to make it easier and possible to educate not only researchers, but also students and the public, as basic knowledge in bioinformatics will be part of the arsenal of any individual who takes part in the moral and ethical discussions that are needed to shape a future that is already here.

EMBnet.journal Editorial Board

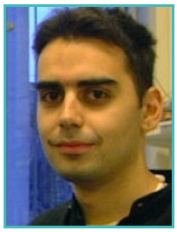
proteinspotlight
> ONE MONTH, ONE PROTEIN <

Protein Spotlight (ISSN 1424-4721) is a periodical electronic review from the SWISS-PROT group of the Swiss Institute of Bioinformatics (SIB). It is published on a monthly basis and consists of articles focused on particular proteins of interest. Each issue is awvailable, free of charge, in HTML or PDF format at http://www.expasy.org/spotlight.
We provide the EMBnet community with a printed version of issue 132. Please let us know if you like this inclusion.

Cover picture: Chinese Lanterns (*Physalis giganteа*), Uppsala, Sweden, 2012. [© Erik Bongcam-Rudloff]

Contents

The future of structural bioinformatics in the post-genomics era; a novel framework to facilitate modern drug design




Dimitrios Vlachakis, Dimosthenis Tsagkrasoulis, Georgia Tsiliki, Sophia Kossida

Bioinformatics & Medical Informatics Laboratory, Biomedical Research Foundation of the Academy of Athens, Athens

Received 2 May 2012; **Published** 15 October 2012

Concept description

A major drawback of *in silico* protein science nowadays is that protein structural comparisons are based on sequence searches. Evolutionary relationships of proteins, protein structure–function predictions and comparative modelling would all benefit from greater use of structural information. There are many examples of protein function annotation where sequence-based searches are insufficient (Dobson *et al.*, 2004). Most RNA viruses, even though they can be evolutionarily linked, share very low sequence identities among their homologous proteins, as they are highly mutagenic. Even though the structures of such are more conserved than their sequences (Illergard *et al.*, 2009), and studies have been carried out in areas such as flexible structural alignment, this fact has nevertheless not yet been satisfactorily utilised (Kolodny *et al.*, 2005; Berbalk *et al.*, 2009; Mayr *et al.*, 2007).

A novel approach that exploits the immense size of genomic databases and links them to structure is presented in this study. Both major types of databases are involved in our methodology: the RCSB-PDB, a database of known biological structures, with information obtained mostly by X-ray crystallography and NMR studies (Rose *et al.*, 2011); and enormous genomic databases, such as the NCBI GenBank and Whole Genome Shotgun (WGS) databases, which contain sequence information from many species (including human) acquired by various large- and small-scale sequencing approaches (Benson *et al.*, 2012; Johnson *et al.*, 2008). At the last count, the PDB contained a total of 77,878 structures, whereas GenBank contains 126,551,501,141 bases in 135,440,924 sequence records, plus another 191,401,393,188 bases in 62,715,288 sequence records in the WGS division.

In our method, PDB structures will need no preliminary analysis, while on the other hand, the DNA sequence data-sets, bigger by several orders of magnitude, will have to undergo special filtering – this will include ruling out low complexity regions and focusing on exonic sequence space, a task that will contribute significant noise-reduction to the initial data. Notably, both major databases involved in this use case have been growing exponentially in size over the last few years (Rose *et al.*, 2011; Benson *et al.*, 2012).

The new methodology will provide the tools required to perform protein similarity searches based on structural rather than sequence information. The input query sequence can either be of known or unknown structure (Figure 1). In each case, the primary amino acid sequence will need to be converted to the amino acid Structural Features Sequence (SFS) format. The SFS format is a novel residue-annotation method based on the structural conformation of each amino acid in the query sequence. For instance, residues forming an α-helix will be replaced with an "H", a β-sheet with an "S", a coil with "C", until all query amino acids have been designated with an SFS value. If the input sequence is of unknown structure, it will be subjected to secondary structure prediction algorithms, and the SFS format will be deduced. The same SFS formatting principle and secondary structure prediction algorithm must be applied to both NCBI databases, which can either be performed on the fly or by the one-off conversion of all known information into a new databank, which will need to be updated regularly. As all entries in the PDB contain secondary structural information, the conversion to SFS for-

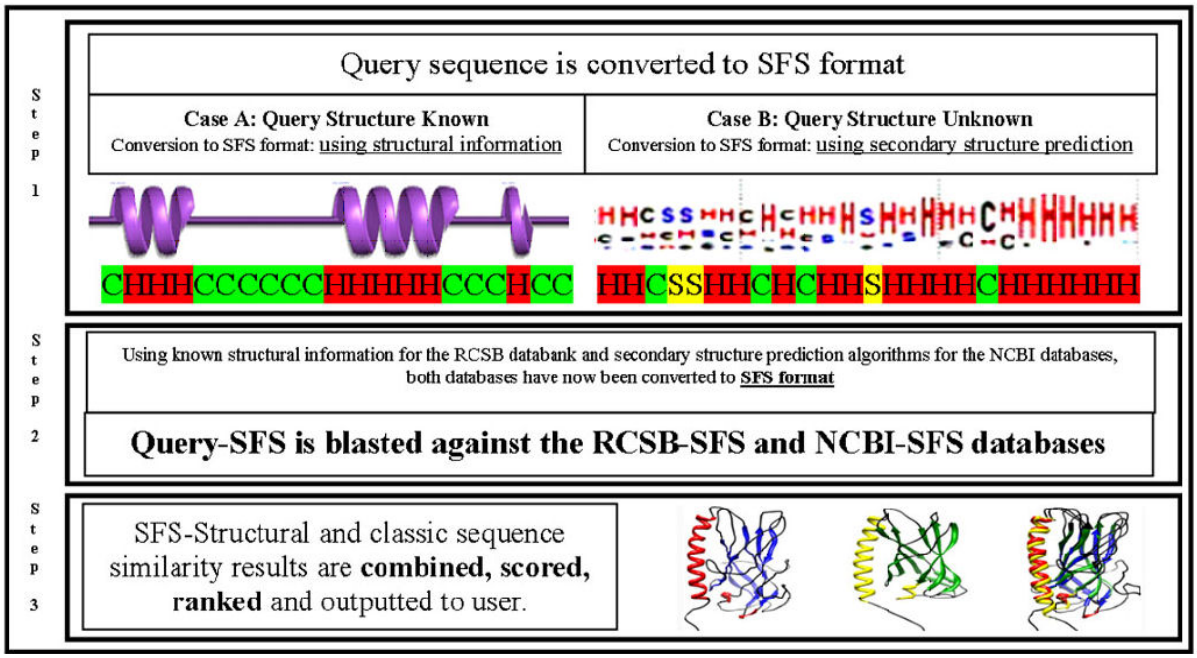


Figure 1. Diagram of the proposed approach.

mat can be performed without any predictions. Our proposed algorithm will be broken down into three parts: in part 1, the query sequence will be converted to SFS format; in part 2, the query-SFS will be structurally aligned against all structures in the PDB-SFS-formatted database and all sequences in the NCBI-SFS-formatted databases; finally, in step 3, structural similarity results will be combined with classic sequence BLAST results and output to the user.

Because our data are, by default, incomplete in the case of genomic sequences that lack structural information, we plan to develop and apply a fast and efficient secondary structure prediction algorithm. However, even upon application of the algorithm, it is still possible to obtain "noisy" data if the prediction score does not clearly indicate structural features. There are two different approaches to deal with this issue. The first is to use multiple secondary structure prediction algorithms, some of which are already established. By applying a variety of different algorithms and approaches on the same sequence string, we will achieve a 'consensus prediction' that will be statistically more reliable. Secondly, we plan to develop a clever algorithm that we will train to recognise and annotate the origin and function of each unknown DNA sequence string using Artificial Intelligence (AI) and machine-learning techniques. Then, by 'homology and comparative approaches', we will be able to 'predict and expect' various structural elements in a given sequence and, accordingly, adjust the weight ratios used by the secondary prediction algorithm. For example, if we obtain noisy/unclear data from the exonic product of a DNA sequence that has been found to contain conserved features of a certain family of transcription factors with α-helical repeats, then the algorithm will 'expect' that sequence to have similar α-helical conformation. It is important to clarify that the 'consensus prediction' and the 'homology and comparative approaches' will only be applied when noisy data appear, saving CPU calculation effort when the data are clean.

Outlook

The real world problem addressed by our new methodology is highly relevant to the general field of biomedicine. Providing a concise and efficient framework for detecting protein structural similarity is bound to be very valuable for experimental drug design. Almost 90% of drugs tested on humans fail owing to unpredicted toxicities. Supplying the bio-pharmaceutical industry with a compendium of easily searchable and retrievable structures against which any substance of interest may be compared in a straight-forward

manner, will enable the filtering out of a significant amount of probable side-effects. This would imply increasing the expected effectiveness of the proposed drug with a simultaneous significant decrease in cost. The pharmaceutical industry would benefit enormously in fields such as drug design and development, by being able to search for similar structural features and active sites for a given drug or inhibitor.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. *Nucleic Acids* Res **40**, D48-53. doi: 10.1093/nar/gkr1202
2. Berbalk C, Schwaiger CS, Lackner P (2009) Accuracy analysis of multiple structure alignments. *Protein Sci* **18**, 2027–2035. doi: 10.1002/pro.213
3. Dobson PD, Cai YD, Stapley BJ, Doig AJ (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr Med Chem* **11**, 2135–2142.
4. Illergard K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* **77**, 499–508. doi: 10.1002/prot.22458
5. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S et al. (2008) NCBI BLAST: a better web interface. *Nucl. Acids Res* **36**, W5-W9. doi: 10.1093/nar/gkn201
6. Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* **346**, 1173–1188. doi: 10.1016/j.jmb.2004.12.032
7. Mayr G, Domingues FS, Lackner P (2007) Comparative analysis of protein structure alignments. *BMC Struct Biol* 7: **50**. doi: 10.1186/1472-6807-7-50
8. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* **39**, D392-401. doi: 10.1093/nar/gkq1021

Speeding up the drug discovery process: structural similarity searches using molecular surfaces

Dimitrios Vlachakis, Georgia Tsiliki, Dimosthenis Tsagkrasoulis, Carla Sofia Carvalho, Vasileios Megalooikonomou, Sofia Kossida

Bioinformatics & Medical Informatics Laboratory, Biomedical Research Foundation of the Academy of Athens, Athens

Received 30 May 2012; **Published** 15 October 2012

Simplifying spatially complicated problems in the field of drug design, pharmacology and 3D molecular modelling is becoming very important, owing to the rapid increase in genomic and structural database sizes. The computational load is immense, and novel innovative approaches are sought, in order to perform comprehensive structural studies and 3D searches at only a fraction of the original time required (Gerld *et al.*, 2011).

Protein docking (PD) and protein-protein interactions (PPI) are two of the most rapidly emerging fields in modern structural bioinformatics. Many studies attempt to justify biological activity and function of small molecules, macromolecules or even molecular complexes using PD and PPI. For example, the majority of the information we have about the molecular processes that take place in the nucleus or the cytoplasm, and affect DNA replication, has been acquired by fast algorithms and machine-learning approaches that investigate protein-protein interactions. Molecular dynamics, genetic and epigenetic networks, systems biology, molecular biology and many other related disciplines use PD and PPI as key research tools. Many databases have been developed in this direction: e.g., the MIPS mammalian protein-protein database, the eF-site molecular surface database, the STRING database of functional protein association networks, BioGRID, VASP, PESDserv and many more (Pagel *et al.*, 2005; Kinoshita and Nakamura, 2003; Szklarczyk et al., 2011; Stark *et al.*, 2011; Chen and Honig, 2010; Das *et al.*, 2010). However, the limitation is that these approaches are modelled simulations using graph-theoretical methods, whose sensitivity and specificity is not always trustworthy. Eventually, human input and insight is required, as the application of current algorithms to all available data is impossible owing to hardware- and time- limitations.

Here, we present a novel strategy to perform similarity searches and molecular docking experiments using protein molecular surfaces. Our approach starts by calculating a series of distinct molecular surfaces for each protein, which are subsequently flattened out, thus reducing 3D information to 2D. Multiple surfaces may be combined to establish 2D Molecular Profile Fingerprints (2DMPFs) unique for each protein. 2DMPFs still retain the original 3D structural information of each protein, and may be analyzed via image-processing and pattern-recognition techniques using sliding windows and similarity-scoring functions. Finally, using fast Fourier transformation algorithms we can move from 2D image data to 1D graph lines, which are unique to each protein and can be used as fingerprints for similarity searches.

The 3D shape, size and surface information of a protein can be depicted using molecular surfaces (Nimrod et al., 2009). There are many different types of molecular surface, the commonest being electrostatic, pocket, lipophilic, b-factor and secondary-element surfaces (Binkowski and Joachimiak, 2008; Yin *et al.*, 2009; Sael *et al.*, 2008, Brylinski and Skolnick, 2010). The first task of our approach will be to calculate a set of fine-grid surfaces of each protein structure available in the RCSB Protein Data Bank (Rose et al., 2011). Then the projection of the protein surfaces from 3D to 2D representation will begin by mapping the molecular surfaces into spherical surfaces of radius proportional to the size of the protein. The resolution block (pixel) is associated with a physical size, and thus has a fixed size for all spherical surfaces. For this step, we will use the SPHARMAT package (Shen and Makedon, 2006). The second step consists of projecting the spherical

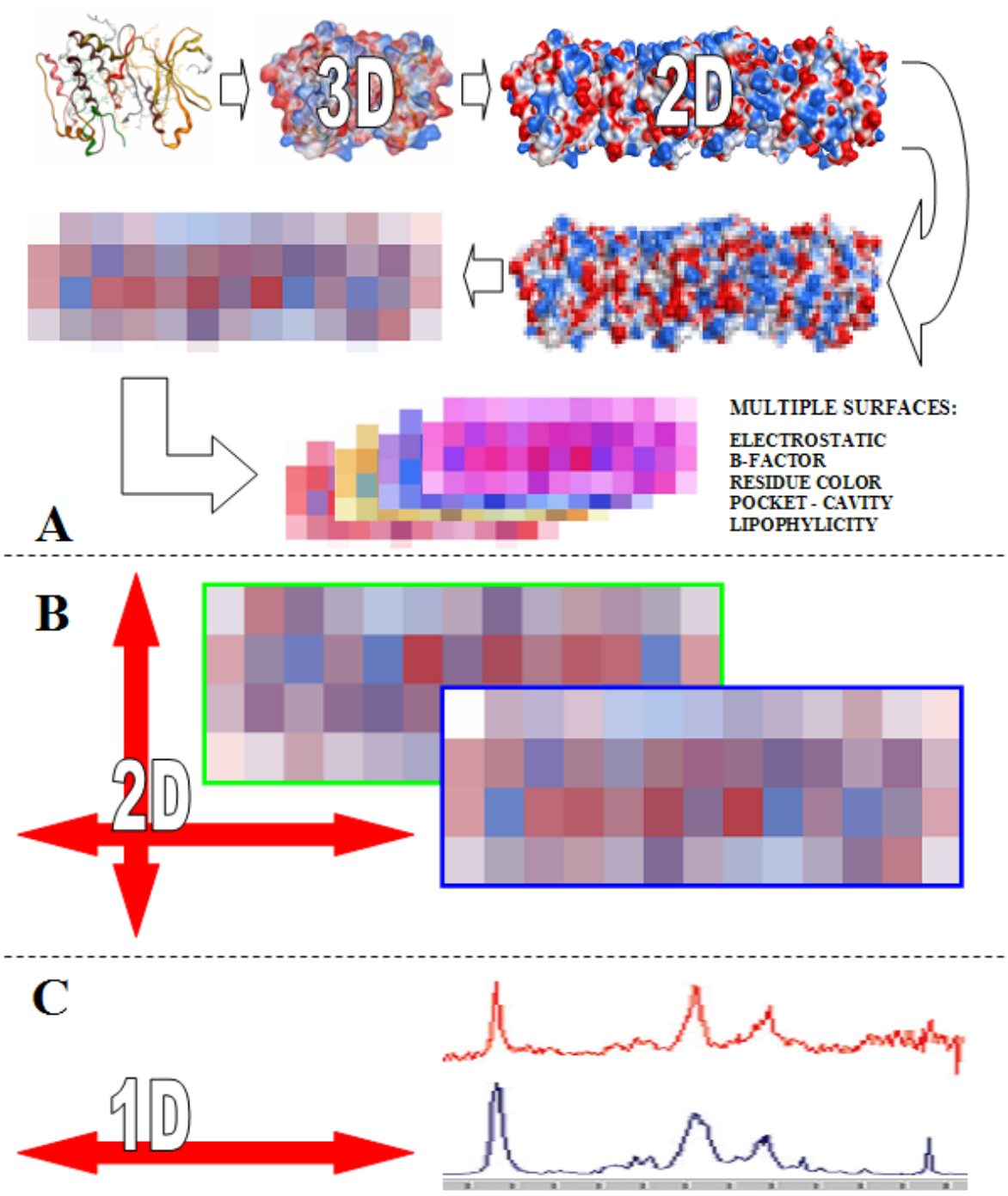


Figure 1. A schematic representation of our proposed approach. (A) The 2D protein profile fingerprint approach. The multiple surface overlapping principle, where various molecular surfaces are combined in a multiple-layer 2D image. (B) The 2D image sliding-surface pattern recognition of either matching or complementary regions. (C) Using fast Fourier transformations we can convert a 2D image to a 1D graph for even faster and more efficient pattern-recognition performance.

surfaces into flat patches along common symmetry axes of the proteins. For this step, we will use the HEALPix package (Gorski et al., 2005). The flat patches will then be the input objects for the measurement of correlations and the search of patterns among proteins. Here, multiple surfaces are being combined (Fig. 1A) and sliding 2D techniques are used for pattern recognition (Fig. 1B).

The actual scanning and filtering of the 2D data for similarity or shape/size complementary patterns will take place in the final step of the algorithm. At this stage, a correlation will be made between the results of the 2D scanning and the biological question. Multiple surfaces will have to be combined using pattern-recognition 2D sliding methodologies. The ultimate objective of our approach is to enable users to explore the computationally demanding 3D conformational space of biomolecular structures using 2D or even 1D data, which will speed up the computational process by reducing data load, without any compromise in protein information. The 1D fingerprint of our 2D images will be obtained by computing the 2-point correlation function of the Fourier transformed 2D images, which still correlate to the original 3D structure. Rather than exploring all the 3D conformational space of large protein structures when performing docking experiments, our approach will be capable of comparing the 2D image fingerprints or 1D Fourier transformed graphs (Fig. 1C) of the given structures, and in a fraction of the original time, returning results that still contain the original 3D structural information.

Multiple studies have been conducted using various correlation measures to identify patterns in 2D data (Xiong and Zhang, 2010). While working well for small datasets, the heterogeneity introduced from increased sample size inevitably reduces the sensitivity and specificity of those approaches. For this reason, we propose a model-based, pattern-recognition algorithm built under a partition-model framework, which is robust against sporadic outliers. Specifically, we assume that each 2D protein profile fingerprint can be presented by an MxN data matrix, where M is the total number of the vertical image resolution blocks and N is the total number of the horizontal ones. 2D data resolution values are categorised based on their individual value range in a scale of $[-\varepsilon, \varepsilon]$: ε is a positive integer empirically derived by simulated data to account for data variability. For each pair of 2D protein fingerprints, we define the difference matrix $Z = \{z_{ij}, i=1, ..., M \text{ and } j=1, ..., N\}$. In this context, zij corresponds to the difference of the values in the corresponding (i,j) cell of the pair of 2D data-files, and $z_{ij} = 0$ if categorical values of the (i,j) cells are identical. The window could slide towards both horizontal and vertical directions, resulting in multiple similarity estimates at each pairwise comparison. The optimal window size will be estimated by minimising the Bayesian Information Criterion (BIC) of the suggested 2-way partition model (Denison *et al.*, 2002). Nested partition models will be also considered. We believe that the multiple overlapping windows solution will allow us to zoom in on the 2D data in a time-inexpensive way, weight their similarities and complementarities by averaging over different neighbourhoods of the data or across data matrices, and also estimate their variability errors. Special care should be given to models' sensitivity to the categorisation scheme and estimation of optimal window size. The suggested approach will be compared with colour similarity metrics along with standard clustering techniques.

In conclusion, our approach introduces a novel technique for searching, evaluating and scoring pattern similarities between a given set of molecular surfaces. Upon calculation of a variety of diverse surface types for each protein, all 3D structural information is converted into a combined, multi-layer 2D image, which can be further simplified to 1D data via Fourier transformation. In this way, we optimise and speed up the time- and CPU-demanding 3D conformational searching, by faster more versatile 2D or 1D datasets, without compromising 3D structural information.

References

1. Binkowski TA and Joachimiak A. (2008) Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. BMC Struct Biol. 8: 45. doi: 10.1186/1472-6807-8-45
2. Brylinski M and Skolnick J. (2010) Comparison of structure- and threading-based approaches to protein functional annotation Proteins. 78, (1), 118–134. doi: 10.1002/prot.22566
3. Chen BY and Honig B (2010) VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. PLoS Comput Biol. 6(8). doi: 10.1371/journal.pcbi.1000881

4. Das S, Krein MP, Breneman CM. (2010) PESDserv: a server for high-throughput comparison of protein binding site surfaces. Bioinformatics. 26(15), 1913–1914. doi: 10.1093/bioinformatics/btq288
5. Denison DGT , Adams NM, Holmes CC, Hand DJ. (2002) Bayesian partition modeling. Comput Stat Data An. 38, (4): 475-485. doi: 10.1016/S0167-9473(01)00073-1
6. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbit PC et al. (2011) The enzyme function initiative. Biochem. 50, 9950-9962. doi: 10.1021/bi201312u
7. Górski KM, Hivon E, Banday AJ, Wandelt BD, Hansen FK et al. (2011) HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. The Astrophysical Journal 622, 759-771. doi: 10.1086/427976
8. Kinoshita K and Nakamura H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci. 12(8), 1589–1595. doi: 10.1110/ps.0368703
9. Shen L, Makedon FS. (2006) Spherical mapping for processing of 3-D closed surfaces. Image and Vision Computing 24, (7):743-761. doi: 10.1016/j.imavis.2006.01.011
10. Nimrod G, Szilágyi A, Leslie C, Ben-Tal N. (2009) Identification of DNA-Binding Proteins Using Structural, Electrostatic and Evolutionary Features. J Mol Biol. 10, 387(4), 1040-1053. doi: 10.1016/j.jmb.2009.02.023
11. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics. 21(6):832-834. doi: 10.1093/bioinformatics/bti115
12. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D et al. (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39, D392-401. doi: 10.1093/nar/gkq1021
13. Sael L, La D, Li B, Rustamov R, Kihara D. (2008) Rapid comparison of properties on protein surface Proteins. 73, (1), 1–10. doi: 10.1002/prot.22141
14. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 39, D698-704. doi: 10.1093/nar/gkq1116
15. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39, D561-568. doi: 10.1093/nar/gkq973
16. Xiong Z and Zhang Y (2010) A critical review of image registration methods Inter J Image Data Fusion. 1, (2): 137-158. doi:10.1080/19479831003802790
17. Yin S, Proctor EA, Lugovskoy AA, Dokholyan NV. (2009) Fast screening of protein surfaces using geometric invariant fingerprints. Proc Natl Acad Sci U S A. 29, 106(39), 16622–16626. doi: 10.1073/pnas.0906146106

EMBnet at ISCB Latin America 2012 Conference on Bioinformatics



Andreas Gisel, Jose Ramon Valverde, Erik Bongcam-Rudloff
CNR - Institute for Biomedical Technologies, Bari, Italy,
Centro Nacional de Biotecnología (CSIC), Madrid, Spain
Institutionen för husdjursgenetik, Uppsala, Sweden

Received 19 July 2012; **Published** 15 October 2012

The second Latin American regional meeting of the International Society for Computational Biology (ISCB-Latin America) took place 17-21 March 2012, in Santiago, Chile. More than 250 people attended, primarily from countries in Latin America. The major aim of ISCB-Latin America 2012[1] was to motivate and inspire young Latin American students and post-docs to conduct the best possible research in the areas of Bioinformatics and Computational Biology.

The first two days of the meeting (17-18 March) were dedicated to hands-on practical tutorials and workshops[2] covering different topics of interest: analysis, comparison and classification of protein structures; genome browsers, with special emphasis in the ENSEMBL system; protein resources and tools; sequence, architecture and protein interactions; algorithms and tools for transcriptomics, and multiple-gene profiling using the open-source platforms R and Bioconductor; immunoinformatics; an introduction to next generation sequencing (NGS) for bioinformaticians; functional genomics and computer-based drug design. EMBnet[3] sponsored two tutorial sessions by financing travel and accommodation for two teachers.

Figure 1. Participants of the ENSEMBL tutorial.

On the first day, Dr. Erik Bongcam-Rudloff (from the Swedish EMBnet node) gave the course, 'Genome browsers, with special emphasis on the ENSEMBL system', offering a broad introduction to biological databases and the ENSEMBL genome browser[4]. More than 15 participants followed this tutorial, using it mainly as an introduction to basic bioinformatics. On the second day, Dr. Andreas Gisel (from the Italian EMBnet node) gave the course 'Next generation sequencing: an introduction for bioinformaticians', a tutorial for more advanced bioinformaticians.

Figure 2. Some of the participants of the NGS tutorial.

With the help of the local administrator, a virtual machine, created by the Italian EMBnet node, including all data for the hands-on, was installed on each classroom computer. In this way, the 35 participants had the same platform, with

1 http://www.iscb.org/iscb-latinamerica2012
2 http://www.iscb.org/iscb-latinamerica2012-program/tutorials
3 http://www.embnet.org
4 http://www.ensembl.org

all the basic NGS-data analysis and visualisation tools. The tutorial included an introduction, NGS-data mapping, and visualisation of mapping data with the Web-based GBrowse tool from the GMOD project[5].

the open source platforms R and Bioconductor'. This session, jointly orga-nised by EMBnet with the Free Software for Life and Health (FreeBIT[6], (CYTED 510RT0391) and the Iberoamerican Society of Bioinformatics (SoIBio[7]), aimed to demonstrate to

Figure 3: Dr. Andreas Gisel explaining NGS-data formats.

In a third session, Dr. José Ramón Valverde (manager of the EMBnet node in Spain) and Dr. J. de las Rivas (from the University of Salamanca, Spain), gave a course entitled, 'Algorithms and tools for transcriptomics, and multiple-gene profiling using and teach 45 students the power and ease-of-use of R and Bioconductor for processing micro-array and transcriptomics data, such as RNA-seq data-sets.

5 http://gmod.org/wiki/GBrowse

6 http://www.free-bit.org) Network of Excellence
7 http://www.soibio.org

ReNaBi-IFB: The French Bioinformatics Infrastructure

Guy Perriere
ReNaBi-IFB: The French Bioinformatics Infrastructure, Pôle Rhône-Alpes de Bioinformatique, Université Claude Bernard, Lyon, France

Received 21 June 2012; **Published** 15 October 2012

In France, the Group of Scientific Interest (GIS), Infrastructures in Biology, Health and Agronomy[1] (IBiSA) is in charge of implementing a concerted policy in terms of infrastructure for life sciences. In the field of bioinformatics, this strategy has resulted in a network of regional platforms (PFs) aimed at fostering the coordination of their activities. At the moment, 13 PFs clustered into six regional centres belong to this network, the ReNaBi[2] (French Bioinformatics Platforms Network), which is also the French national node for EMBnet. Those six regional centres span the French territory (ReNaBi-NE, North-East; PRABI, Rhône-Alpes region; ReNaBi-GS, Great South; ReNaBi-SO, South-West; ReNaBi-GO, Great West; APLIBIO, Paris area), and they are all embedded in bioinformatics research laboratories. Their corresponding manpower is about 100 Full-Time Equivalents (FTEs) in terms of permanent staff, and 57 FTEs for people hired on fixed-term contracts; this represents about 30% of the whole French bioinformatics community.

The main limitation of this network is that the resources and know-how are geographically distributed and somehow redundant. This makes the pooling of resources and expertise more difficult to harness. In addition, this 'scattered' structure is not intelligible from outside the French bioinformatics community, particularly for international partners. Therefore, the ReNaBi is moving toward a more centralised structure, the French Bioinformatics Infrastructure (IFB). This will be based on:

- **a national node** (IFB-core), having its own head, staff and IT infrastructure. Its role will be to serve as the unique entry point for requests of services from the biological community, to coordinate and structure the activities of the PFs and to ensure consistent coordination between IFB and national users (in particular, other large national infrastructures producing 'omics' data);
- existing **regional PFs**, where the methodological and user-training know-how is to be found, and that currently provide support to projects with biologists in their respective regions. PFs will be structured more assertively around thematic poles characterised by their international visibility and/or biological data specificity.

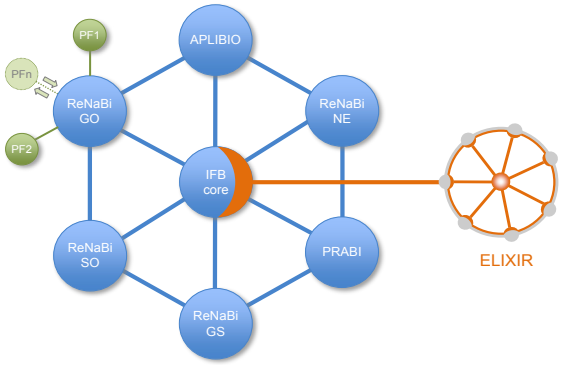


Figure 1. IFB structure and its link with ELIXIR. The six regional centres are linked to the IFB-core, and each centre integrates a variable number of PFs. Integration or removal of a PF is under the responsibility of the regional centre scientific board.

In order to fulfill its missions to the French biology and biomedical research community and, in particular, to ensure that IFB will address the proper analysis and data-management challenges, the IFB operation mode will be 'project-oriented'. IFB will provide development, maintenance and training support for clearly defined biological, biomedical or technological projects, based on a transparent reviewing process to ensure the highest quality in project selection. The targeted projects will fall within several categories depending upon their size or strategic status:

- large-scale institutional projects and projects with other national infrastructures;
- biology and biomedical research projects;
- services offered to industry;
- technological projects.

IFB, through the IFB-core, will also become the French node of ELIXIR[3]. IFB intends to address, at the national level, issues similar to those that ELIXIR

1 http://www.ibisa.net/
2 http://www.renabi.fr/
3 http://www.elixir-europe.org/

aims to tackle: data and computer infrastructure, tools and standards, training. IFB structuring will allow the mobilisation of resources at the national scale, not only to develop or enhance all the components that meet ELIXIR demands for excellence, but also to ensure the sustainability of these components.

Finally, an important task of IFB will involve monitoring the PFs' activities, collecting associated data and figures to help the supervisory authorities to get a more precise overview of the bioinformatics landscape. IFB, taking advantage of the gathering of PFs' members around thematic poles, will be in charge of commissioning the publication of 'white papers' about bioinformatics issues, especially those regarding the demands of other national infrastructures. It is therefore expected that IFB will significantly contribute to prospective reflection in the field of bioinformatics.

Minimum Information About a Peptide Array Experiment (MIAPepAE)


Gordon Botha, Judit Kumuthini*
Centre for Proteomic and Genomic Research (CPGR), Cape Town, South Africa

* corresponding author (jkumuthini@gmail.com)

Submited 17 November 2011; **Accepted** 29 June 2012; **Published** 15 October 2012
Competing Interests: none

Abstract

Peptide-array screening is currently a well-established high-throughput technique with growing numbers of applications. Peptide-array technology is used for protein recognition, quantification of peptide expression levels, and detection of protein-protein interactions. The use of protein/peptide arrays in medical life science studies is becoming increasingly widespread. Their increased use in diagnostic applications and protein function profiling calls for a standardised set of guidelines to be followed by future experimenters to enable reproducible, high-quality data and accurate findings.

We aim to provide preliminary guidelines describing the Minimum Information About a Peptide-Array Experiment (MIAPepAE). We propose a checklist of data and meta-data that should accompany a peptide-array experiment, and invite fellow researchers in the field to collaborate in this effort to create a sustainable and coherent set of guidelines for the benefit of the protein/peptide-array research community. Although this article focuses on spotting peptide arrays, MIAPepAE is intended to be a work-in-progress to be adopted for other peptide-array types, such as in situ synthesised peptide arrays.

Introduction

The use of protein/peptide arrays in medical life science studies is becoming increasingly widespread (Reimer et al., 2002; Wulfkuhle et al., 2003; Cretich *et al.*, 2006). Broadly speaking, they are used for two main purposes: diagnostic applications (bio-markers or antibody detection) and protein function profiling. Peptide arrays are powerful diagnostic tools, as they allow both multiple analyses of identical samples and single-instance analyses of differential samples. For example, they have been applied to immune-response profiling experiments by measuring antibody-antigen interactions (Davies et al., 2005; Ingvarsson *et al.*, 2008; Andresen and Grötzinger, 2009); they have also been instrumental in protein-function profiling studies (Katz et al., 2011), in part because they use very little sample material and can process many proteins in parallel (Haab, 2001), and partly also because they can quantify very low concentrations of protein (Korf *et al.*, 2008) and take into account protein/peptide tertiary structures. Overall, peptide arrays are becoming pivotal to protein studies, spurring developments in related fields.

The technology and methodology is steadily advancing, in terms of slide preparation (Kopf et al., 2005; Beyer *et al.*, 2006) and sample preparation (Ghazani *et al.*, 2006; Usui *et al.*, 2006), and, in turn, is leading bioinformaticians to develop new software tools (Li *et al.*, 2005) and Web applications (Li *et al.*, 2009). Specific statistical techniques for the analysis of peptide arrays have also been developed (Royce *et al.*, 2006). Furthermore, high-throughput sequencing methods, such as real time PCR (Heid *et al.*, 1996), have delivered

an abundance of genomic and proteomic data for many species (Love *et al.*, 1990; Blattner *et al.*, 1997; Dean *et al.*, 2002). With so much proteomic information and analysis tools available, it is inevitable that many more peptide-array experiments will be conducted in the foreseeable future.

Our aim is to provide preliminary guidelines for the Minimum Information About a Peptide-Array Experiment (MIAPepAE). We propose a checklist of data and meta-data that should accompany a peptide-array experiment, aiming to fulfill the following main objectives:

• MIAPepAE should provide authors, reviewers, editors and readers with the specifics required to critically evaluate, understand and reproduce a peptide-array experiment;

• MIAPepAE should provide sufficient information to aggregate/integrate similar experimental data, independently of the platform on which the experiment was performed;

• MIAPepAE should allow secondary data, such as clinical patient and epidemiology data, to be integrated, enabling the extraction of more meaningful information from peptide-array experiments.

We emphasise meta-data pertaining to the sample. Variation in preparation of protein/peptide samples and their assaying to the array slides can be a major contributor to experimental variation and, as such, warrants a focused effort toward the proposed guidelines.

In the interest of coherent and coordinated development of such guidelines, the project is registered on the MIBBI porta[1]. The MIBBI project is a collaboration between leaders in the biological and biomedical fields, acting as a meeting point for the coordination of minimum information guidelines and checklists (Taylor *et al.*, 2008)

We have also based our checklist format on the guidelines for peptide-array experiments provided in the Minimum Information About a Proteomics Experiment (MIAPE) article (Taylor *et al.*, 2007), and the Minimum Information About a Microarray Experiment (MIAME) article (Brazma *et al.*, 2001b). The original MIAME checklist for microarray experiments has been revised (see (Abeygunawardena, 2007)), and we have based our checklist for peptide-arrays on the revised checklist. Hence, we have drafted our checklist with the following main subjects: Raw Data, Final Processed Data for Set of Hybridisations, Sample Annotation and Experimental Factors, Experimental Design, Sufficient Annotation of Array Design, Essential Experimental and Data-Processing Protocols.

We endeavour to adhere to two criteria introduced by the MIAPE article: those of *Sufficiency* and *Practicability*. *Sufficiency* states that the minimum information requirements are constructed in such a way that the reviewer is able to "understand and critically evaluate the interpretation and conclusions". The reader must also be able to support the findings. *Practicability* states that the incorporation of a minimum information requirement for a proteomic experiment need not be so taxing on the experimenters that its adoption is impaired.

The checklist is still under development and will undoubtedly undergo revision as more peptide array experiments are performed and more comments and suggestions from colleagues in the field are incorporated.

Key Concepts

Our approach towards the formulation of guidelines for a peptide-array experiment takes several key concepts into account. These need to be defined clearly before proceeding, as this is necessary to interpret our guidelines.

Microarray Nomenclature

We have compiled a nomenclature from previous definitions (Brazma et al., 2001a; Royce *et al.*, 2006). The molecules bonded to the slide at the time of manufacture are termed *probes*. Any subsequent binding molecules are termed targets. A *spot* or *feature* is defined as a group of probes with identical sequences, concentrated at a known position on the microarray. A group of *targets* from the same biological entity is defined as a *sample*. One instance of the introduction of one or more samples to the array is known as *probing*. Finally, a series of probing to investigate a hypothesis is known as an *experiment*.

Unique Peptide

A unique peptide, as used in a peptide-array experiment, should conform to the following

1 http://mibbi.org/index.php/MIBBI_portal

properties. It should have a unique identification number (ID) such as a National Centre for Biotechnology Information (NCBI) or a Protein Data Bank (PDB) number. If the peptide is synthetic, the full amino-acid sequence must be made available. A list of the protein(s) in which the peptide can be found should be given, including the starting position in the protein. The peptide length should be specified, and the overlap used when aligning the peptide to a protein. Finally, any unidentified/ambiguous amino acids within the peptide sequence must be noted.

Table 1. MIAPepAE checklist for authors, revie-wers and editors. All essential information (E) must be submitted with the manuscript. Desirable information (D) should be submitted if available.

EXPERIMENTER INFO			
Author (submitter), laboratory, contact information (e-mail, postal address), links (URL), citation			
RAW DATA			
Typically, these are the data-files produced by microarray image-analysis software			
		IMPORTANCE	**CHECKLIST**
Raw data-files provided			
	Native format	E	
	Type: *e.g.*, image, binary data	D	
	The file matches the respective array design	D	
	Scanned image files for each slide	D	
	Data location	E	
FINAL PROCESSED DATA FOR SET OF HYBRIDISATIONS (EXPERIMENT)			
Normalised/Summarised data on which conclusions are based			
		IMPORTANCE	**CHECKLIST**
Processed (normalised) data-files		E	
Normalisation application: *e.g.*, pin-to-pin, array-to-array, slide-to-slide, background correction		E	
Normalisation method		E	
The identifiers match the array annotation/ location		D	
Control(s) on which normalisation was based		E	
SAMPLE ANNOTATION & EXPERIMENTAL FACTORS			
Describes the key experimental variables in the experiment. Additional information regarding sample, such as storage conditions, preparation methods, etc., are of great importance.			
		IMPORTANCE	**CHECKLIST**
Basic experimental factors (dose, time, disease state, treatment) provided for all samples		E	
Additional sample information			
	Sample type	D	
	Sample storage condition	D	
	Sample dilution buffer	D	

	Sample name/annotation	E	
	Sample dilution used in the assay	E	
	Blocking agent	D	
	Detection antibody	E	
	Concentration of detection antibody	E	
	Hybridisation and washing conditions	D	
	Type of dye	D	
	Source organism (NCBI ta-xonomy)	D	
	Laboratory protocol for sample treatment (name, version, availability)	D	
Any post-printing processing, including cross-linking			
	Protein from which peptide was extracted (incl. ID) (NCBI/ UniProtKB/ SwissProt)	E	
	Peptide position in protein	E	
	Peptide overlap in protein alignment	E	
	Peptide conservancy	E	
	Peptide/protein sequence ID (NCBI/ UniProtKB /SwissProt)	E	

EXPERIMENTAL DESIGN

Describes the basic way in which the experiment was set up. Associations between samples and raw data generated from using these samples are critical. Note that the representation of an experimental design is best done via a graphical representation. The MAGE-TAB spreadsheet template (see text) provides a simple format for encoding such graphs.

		IMPORTANCE	CHECKLIST
Experimental design description			
	Table showing (sample) - (raw-data file) associations	E	
	Essential relationships between sample and array biomaterial noted	E	
	Experiment variables: *e.g.*, treated vs untreated	E	
Replicates			
	Identify which, if any, of the arrays are replicates	E	
	Identify whether replicates are technical/biological	E	

SUFFICIENT ANNOTATION OF ARRAY DESIGN			
Essential information regarding array design, such as layout, probe information, slide surface preparation, etc.			
		IMPORTANCE	**CHECKLIST**
Probe sequence information			
	Probe sequence database ID or complete peptide sequence, if synthetic for every probe**	D	
	**: Disclosure of the probe sequence is highly desirable and strongly encouraged. However, as not all commercial predesigned assay vendors provide this information, it cannot be an essential requirement. Use of such assays is advised against.		
Controls			
	Positive controls, incl. sequence	E	
	Negative controls, incl. sequence	E	
	Synthetic/organic	E	
	Other buffer or empty spots?	E	
Array Design			
	GenePix Array List GAL file (or similar) with complete grid and labelling for all probes on array (incl. replicates, controls, sequence, and annotation if possible)	E	
	Surface type	D	
	Number of pins per array	E	
Slide Preparation			
	Number of array per slide	E	
	Preparation info (blocked, *etc.*)	D	

ESSENTIAL EXPERIMENTAL AND DATA PROCESSING PROTOCOLS			
Essential experimental and data-processing protocols are typically described in the methodology/method. If protocols that allow for variable/user-defined variables are used, these must be adequately described. As for novel analysis methods, the protocol should be sufficiently documented to allow a reviewer to fully understand the process involved. Most software packages are able to output these parameter settings into files such as ArrayPro 'Spot Descriptor' or ArrayPro 'Grid Overlay' files (APG).			
		IMPORTANCE	**CHECKLIST**
Spot intensities			
	Method (cell boundary definition/edge detection, *etc.*)	D	

	Pixels per spot	D	
	Spot dimensions (approximate diameter)	D	
	Spot local background dimensions (approximate diameter)		
	Net intensity calculation (raw minus mean background, raw minus spot background, *etc.*)	E	
Grid-finding methodology			
	Grid layout file (incl. spacing between sub-grids, grid rotation, spot shape and size, *etc.*)	E	
Background intensities			
	Method (local ring, local corners, global from image, global from background cells, *etc.*)	E	
Normalisation			
	Method (Loess, quantile, scaling, *etc.*)	E	
	Normalisation parameter (mean, median, *etc.*)	E	
	Spots used for normalisation (controls, all, subset, *etc.*)	E	
Instruments used			
	Scanner name	D	
	Model	D	
	Proprietary software name, version	D	
Data-extraction software used			
	Name	E	
	Version	D	
	Gain setting	E	
	Minimum threshold	E	
	Macro or script used for data extraction	D	
	Settings file	E	
Data Filtration method			
	Negative controls	D	
	Signal qualities (…from PROCAT)	D	
	Flagged spots criteria	E	
	Criteria 1	E	
	Criteria 2, *etc.*	E	

Reproducibility

Our guidelines aim to maximise the reproducibility of an experiment. They will also ease the interpretation of findings by peers, as a clear idea of experimental procedures will more effectively orient a reviewer.

Comparability and Re-usability

Another key concept that we want to capture in the guidelines is that of platform-independent comparability. Findings between studies can be compared effectively if standardised data formats are in place. Furthermore, if data are extracted in a concise and correct manner, they can be used in subsequent experiments. We feel that quality of data supersedes quantity of data, and using a concise method of data extraction from peptide arrays can greatly increase the experimenter's ability to sort biological meaning from experimental error.

Specificity

The specificity of an experiment, or an experiment in a more general case, measures the ability to correctly classify positive events. In a peptide-array context, this could measure the probability that a peptide/protein-binding event is in fact a specific binding, and hence biologically significant, and not due to a non-specific binding event or experimental error.

Quantification

The crux of the guidelines is to enable correct quantification of spot intensity within an array. It is paramount that spot intensities are biologically significant readings and not the result of experimental variation. The correct quantification of experimental parameters lends itself to effective verification of findings. We aim to achieve this with the proposed guidelines.

Conclusion

We have provided a checklist for capturing essential information when conducting peptide-array experiments. By conforming to this checklist, experimenters will:

provide authors, reviewers, editors and readers with the specifics required to critically evaluate, understand and reproduce a peptide-array experiment. This will lead to more accurate conclusions and higher quality data;

be able to compare and combine experiments across different platforms, greatly enhancing the re-usability of the data;

be able to extract and combine meta-data from experiments that might bring to light interesting observations. In so doing, experimental data can be utilised fully to discover biologically relevant observations.

The MIAPepAE form/format contains most of the required fields and sections in one document type, and allows for continuous updating as procedural standards become apparent from discussions within the community. Certainly, as technological advancements are made, the guideline will be appropriately adjusted. The document is version controlled and is available on the MIBBI portal.

In the interests of speeding up adoption of the MIAPepAE checklist in peptide-array experiments, we urge experimenters to provide at least the essential fields in an electronic format with published data and articles. Only with other researchers' input can the ease of conforming to the standards, and accuracy of field prioritisation within the checklist, be assessed. We do, however, note that, for the full benefits of the MIAPepAE guidelines to be reached, project conformity will have to be enforced at a higher level. Like other minimum information protocols, compliance can be required for: i) the publication of research articles (at journal level); ii) data submission to proprietary and public data repositories (at project and framework level); iii) funding and grant proposals (from funders); and possibly iv) encouragement from open-source project repositories.

We hope that the MIAPepAE guidelines are useful to data generators, data consumers and end users. This will, however, depend entirely on the willingness of the scientific community to adopt the guidelines and, more importantly, the willingness of fellow peptide-array experimenters to contribute to (and criticise) the development of the guidelines. In the end, the success of this project depends entirely on the community that it serves.

References

1. Andresen H, and Grötzinger C (2009). Deciphering the antibodyome - peptide arrays for serum antibody biomarker diagnostics. *Current Proteomics* **6 (1)**, 1-12.

2. Beyer M, Felgenhauer T, Ralf Bischoff F, Breitling F, and Stadler V (2006). A novel glass slide-based peptide array support with high functionality resisting non-specific protein adsorption. *Biomaterials* **27**, 3505–3514.
3. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland *et al.* (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science* **277**, 1453-1462.
4. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman et al. (2001a). Minimum information about a microarray experiment (MIAME) toward standards for microarray data. *Nat Genet* **29**, 365-371.
5. Cretich M, Damin F, Pirri G and Chiari M (2006). Protein and peptide arrays: Recent trends and new directions. *Biomolecular Engineering* **23**, 77-88.
6. Davies DH, Liang X, Hernandez JE, Randall A, Hirst S et al. (2005). Profiling the humoral immune response to infection by using proteome microarrays: High-throughput vaccine and diagnostic antigen discovery. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 547-552.
7. Dean FB, Hosono S, Fang L, Wu X, Faruqi *et al.* (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5261-5266.
8. Ghazani AA, Lee JA, Klostranec J, Xiang Q, Dacosta *et al.* (2006). High Throughput Quantification of Protein Expression of Cancer Antigens in Tissue Microarray Using Quantum Dot Nanocrystals. *Nano Letters* **6**, 2881-2886.
9. Haab BB (2001). Advances in protein microarray technology for protein expression and interaction profiling. *Curr Opin Drug Discov Devel* **4**, 116-123.
10. Heid CA, Stevens J, Livak KJ and Williams PM (1996). Real time quantitative PCR. *Genome Research* **6**, 986-994.
11. Ingvarsson J, Wingren C, Carlsson A, Ellmark P, Wahren *et al.* (2008). Detection of pancreatic cancer using antibody microarray-based serum protein profiling. *Proteomics* **8**, 2211–2219.
12. Katz C, Levy-Beladev L, Rotem-Bamberger S, Rito T, Rüdiger SGD *et al.* (2011). Studying protein–protein interactions using peptide arrays. *Chem. Soc. Rev.* **40**, 2131-2145
13. Kopf E, Shnitzer D and Zharhary D (2005). Panorama Ab Microarray Cell Signaling kit: a unique tool for protein expression analysis. *Proteomics* **5**, 2412–2416.
14. Korf U, Derdak S, Tresch A, Henjes F, Schumacher et al. (2008). Quantitative protein microarrays for time-resolved measurements of protein phosphorylation. *Proteomics* **8**, 4603–4612.
15. Li T, Zuo Z, Zhu Q. Hong A, Zhou X and Gao X (2009). Web-based design of peptide microarrays using microPepArray Pro. *Methods Mol. Biol* **570**, 391–401.
16. Li X, Yi EC, Kemp CJ, Zhang H and Aebersold R (2005). A Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry. *Molecular & Cellular Proteomics* **4**, 1328 –1340.
17. Love JM, Knight AM, McAleer MA and Todd JA (1990). Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. *Nucleic Acids Research* **18**, 4123-4130.
18. Niran Abeygunawardena (2007). MIAME 2.0 - MIAME – FGED http://www.mged.org/Workgroups/MIAME/miame.html.
19. Reimer U, Reineke U and Schneider-Mergener J (2002). Peptide arrays: from macro to micro. *Current Opinion in Biotechnology* **13**, 315–320.
20. Royce TE, Rozowsky JS, Luscombe NM, Emanuelsson O, Yu *et al.* (2006). Extrapolating traditional DNA microarray statistics to tiling and protein microarray technologies. *Meth. Enzymol* **411**, 282-311.
21. Sakanyan V (2005). High-throughput and multiplexed protein array technology: protein-DNA and protein-protein interactions. *Journal of Chromatography B* **815**, **77–95**.
22. Stears RL, Martinsky T and Schena M (2003). Trends in microarray analysis. *Nat Med* **9**, 140–145.
23. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R *et al.* (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotech* **26**, 889–896.
24. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian *et al.* (2007). The minimum information about a proteomics experiment (MIAPE). *Nat Biotech* **25**, 887–893.
25. Usui K, Tomizaki K, Ohyama T, Nokihara K and Mihara H (2006). A novel peptide microarray for protein detection and analysis utilizing a dry peptide array system. *Mol. BioSyst.* **2**, 113.
26. Wulfkuhle JD, Aquino JA, Calvert VS, Fishman DA, Coukos G *et al.* (2003). Signal pathway profiling of ovarian cancer from human tissue specimens using reverse-phase protein microarrays. *Proteomics* **3**, 2085-2090.

Using ARC-based grids for NGS read mapping – Grid interface for BWA

Kimmo Mattila*
CSC - IT Center For Science, Espoo, Finland

* corresponding author (Kimmo.Mattila@csc.fi)

Received 3 February 2012; **Accepted** 2 May 2012; **Published** 15 October 2012
Competing Interests: none

Abstract

This technical note describes a command-line Grid interface for the Burrows-Wheeler Aligner (BWA) read-mapping program. With this interface, BWA jobs can easily utilise Advanced Resource Connector (ARC) middleware-based Grid environments. The interface automatically splits the mapping task into sub-tasks that are executed in parallel in a computing Grid. This approach can significantly speed up the read-mapping process.

Availability: **http://www.csc.fi/english/research/sciences/bioscience/programs/BWA**

Introduction

Aligning large sets of short sequences, reads, to a reference sequence set is an essential step in many Next Generation Sequencing (NGS)-based analysis workflows. At the moment, several read-mapping programs are available to perform this alignment task: e.g., tools like Bowtie (Langmead et al, 2009), Burrows-Wheeler Aligner (BWA) (Li and Durbim, 2009) and Maq (Li *et al.* 2008). The resulting alignment files are further analysed with different tools depending on the case.

Different mapping tools apply different algorithms and heuristics to do the alignment. However, the common trend seems to be that tools that produce higher mapping accuracy also require more computing power. Even though the mapping tools are extremely fast compared to the previous generation of sequence alignment tools, mapping hundreds of millions of reads against the reference genome can require weeks of computing time on a normal desktop computer.

One way to speed up the mapping tasks, in addition to parallel computing or special hardware, is to use Grid computing. The read-mapping tasks suit well to Grid computing, as the analysis task can normally be divided into numerous sub-tasks that can later be merged back together. In this note, we present a command-line Grid interface for the commonly used BWA. The goal of the interface is to allow Linux users to utilise Advanced Resource Connector (ARC) (Ellert *et al*, 2007) – middleware-based computing Grids for NGS mapping jobs – without any knowledge about the Grid middleware itself.

The Grid interface of BWA is currently in use at the servers of CSC, IT Center for Science, but in principle it could be installed on any Linux server. Similar interfaces are also available for other tools, like the SHRiMP read-mapping tool (David M et al. 2011).

Methodology

The BWA Grid submission command, grid _ bwa, executes the following five basic steps: 1) checking input; 2) indexing the reference; 3) splitting the mapping task into sub-jobs; 4) executing sub-jobs in the Grid; 5) collecting the results. These steps are discussed more in detail below.

1. Checking the environment, input data and parameters

In the beginning, the *grid_bwa* command checks that the input files have the expected file types (fasta for the reference genome, fastq for the read files). In the case of paired-end analysis, the two query files are checked to have equal amounts of sequences. Even though checking the number of sequences in a file is a simple task, it may take tens of minutes if the input files contain hundreds of millions of sequences.

2. Indexing the reference and storing the indexes to a central repository

Indexing the reference genome or sequence set is the first step of the BWA analysis. In normal usage, this indexing is done with a separate command that launches the actual mapping task. In the *grid_bwa* command, the indexing is integrated as an automatic part of the alignment command. Further, the *grid_bwa* first checks if indices for the reference genome already exist in the user's personal data repository (storage element) in the Grid environment. If indices are not found, they are computed and stored in the Grid repository.

3. Splitting the mapping task into sub-jobs

Splitting the jobs is straightforward, but for large input files it can take several hours. The resulting query subsets are copied to sub-job-specific temporary directories, after which a grid job description file and corresponding job execution file are generated for each sub-job directory. The *grid job description files* use the ARC XRSL format. A *job-description file* contains information about input files that need to be copied to the remote execution site and the output files that should be retrieved when the job is finished. The job-description file also contains the execution time, memory and software requirements of the job. In the case of BWA jobs, fixed 24h and 8GB reservations are used.

The *job execution file* is a shell script that contains commands that are needed to: 1) unpack the input data; 2) run the actual mapping task; 3) post-process the mapping results.

In principle, large mapping tasks could be split into millions of sub-jobs. In practice, splitting the task into more than a few thousands of sub-jobs is not feasible, because managing large amounts or sub-directories is inefficient. Further, the optimal execution time for sub-jobs is in the range of a few hours. In the case of very short sub-jobs, the overhead caused by the job pre- and post-processing can become relatively large. On the other hand, very long sub-jobs often have to wait longer in the batch queues, which again increases the throughput time of the jobs. By default, the command splits the job into about 300 sub-jobs. The number of sub-jobs can be modified with the option *-nsplit.* For small jobs, the job-splitting and result-merging steps can be skipped by setting *-nsplit 1.*

4. Executing the sub-jobs in grid

When the job-splitting phase is ready, the job-specific temporary directory contains from tens to a few thousands of sub-directories, each containing all the data for one ARC middleware-based Grid job. *grid_bwa* automatically checks which computing resources (*i.e.*, clusters connected to the Grid) are available for the user. To submit the jobs to the remote clusters, a job-manager tool, written at CSC, is used. This job manager tries to optimise the usage of the Grid environment. It follows how many jobs are queueing in the clusters, and sends more jobs only when there are free resources available. The job manager keeps track of the executed sub-jobs and starts sending more jobs to those clusters that execute the jobs most efficiently. As some of the clusters may not work properly, part of the jobs may fail for technical reasons. If this happens, the failed sub-jobs are resubmitted to other clusters three times before they are considered as 'failed' sub-jobs. When a job finishes successfully, the job manager retrieves the result files from the Grid to the sub-job-specific directory at the local computer. In the beginning, only a few jobs run, but gradually more and more jobs get to the execution phase and, after a while, there can be hundreds of BWA tasks being executed at the same time, depending on the amount of suitable Grid resources.

5. Collecting the results

When all the sub-jobs are ready, the alignment files are merged into one indexed bam file using the SAMtools package (Li *et al.* 2009). The query sequences from jobs that have not produced a result file are collected into a separate file. For example, if some query subsets require exceptionally long execution time, they may fail because they exceeded the computing-time limit. Such failed query sets can be processed by resubmitting them with the *grid_bwa* command. During

this second iteration, the number of query sequences in each sub-job is normally so small that all the sub-jobs get processed quickly enough.

Performance

Evaluation of the performance of Grid based tools is difficult. In principle, the more Grid resources you have available, the more sub-jobs you can execute in the same time and the faster all the sub-jobs are finished. However, as the general work-load in the Grid environments varies from time to time, so does the computing power that the Grid job can utilise. Pre- and post-processing can also take some time and, because of this, small alignment tasks that do not require more than few hours of computing time do not actually benefit from splitting the job. For larger jobs, utilising distributed Grid computing enables running in one night tasks that would take weeks on a normal desktop computer. Table 1 shows statistics for a *grid _ bwa* run, where 264 million read pairs (2*101 bp) were mapped against a pre-indexed human genome, using the default paired-end mapping parameters of BWA. *grid _ bwa* split the task into 301 sub-jobs that were processed with a small test cluster (based on 2.67 GHz Intel Xeon CPUs). The average execution time for one sub-job, executed with six computing cores, was about 40 minutes. The environment used in this test was able to process 16 sub-jobs simultaneously (reserving a total of 96 computing cores). With these resources, the mapping task was executed in 14 hours. In comparison, running the same analysis as one job using six cores of an 2.26 GHz Intel Xeon X7560-based server takes about 34 hours (including the conversion of the results into indexed bam files). Thus, in many cases, the actual speed-up gained by using the Grid interface is only moderate. However, in this comparison, we are ignoring the time required to copy the input data to the computing cluster and the time that the job has to wait in the batch queue before the execution starts.

Step	Duration
Checking input and parameters	36 min
Checking pre-calculated indexes	1 min
Splitting the job into 301 subjobs	2h 11 min
Executing the 301 sub-jobs in the grid	5h 43 min
Merging results	4h 26 min
Total	**13h 57 min**

Table 1. Wall-clock times used by the different steps of a grid _ bwa run. In the sample task, 264 million reads pairs (read length = 101 bp) were mapped against a pre-indexed human genome using BWA default parameter.

Command line interface

The BWA Grid-submission command, *grid _ bwa*, is designed to look much like the normal BWA command. The Grid related tasks are all integrated inside the command-line interface, so it is enough for a user just to create the Grid-proxy certificate on the client machine, before executing the job-submission command. Typically, the certificate is generated with ARC command *arc proxy*.

In normal BWA usage, the basic command to map reads in file *reads.fq* to reference genome *genome.fa*, is:

```
bwa aln genome.fa reads.fq > aln _ sa.sai
```

With grid _ bwa, the same task could be executed as a distributed Grid job with command:

```
grid _ bwa aln -query reads.fq -ref
genome.fa -out aln.bam
```

The two major differences between the normal BWA command and the Grid version is that: 1) in the Grid version, the query, reference and output file must be defined with explicit options (*-query*, *-ref* and *-out*); 2) the .sai formatted BWA output files are automatically converted into bam format using *bwa samse* or *bwa sampe* commands and the SAMtools package.

All other BWA parameters can be defined with normal command-line options. For example, adjusting seed length to 24, and defining bar-code

```
grid _ bwa aln -quey genome.fa -ref reads.
fq -out aln.bam -l 24 -B 6
```

In the case of paired-end data, *grid _ bwa* has yet another difference. Normally, paired-end alignment is computed using two *bwa aln* commands, from which the results are combined with the *bwa sampe* command. For example:

```
bwa aln genome.fa reads1.fq > aln1.sai
bwa aln genome.fa reads2.fq > aln2.sai
bwa sampe genome.fa aln1.sai aln2.sai
reads1.fq reads2.fq > aln.sam
```

In the case of *grid _ bwa*, all these steps are executed using just a single command:

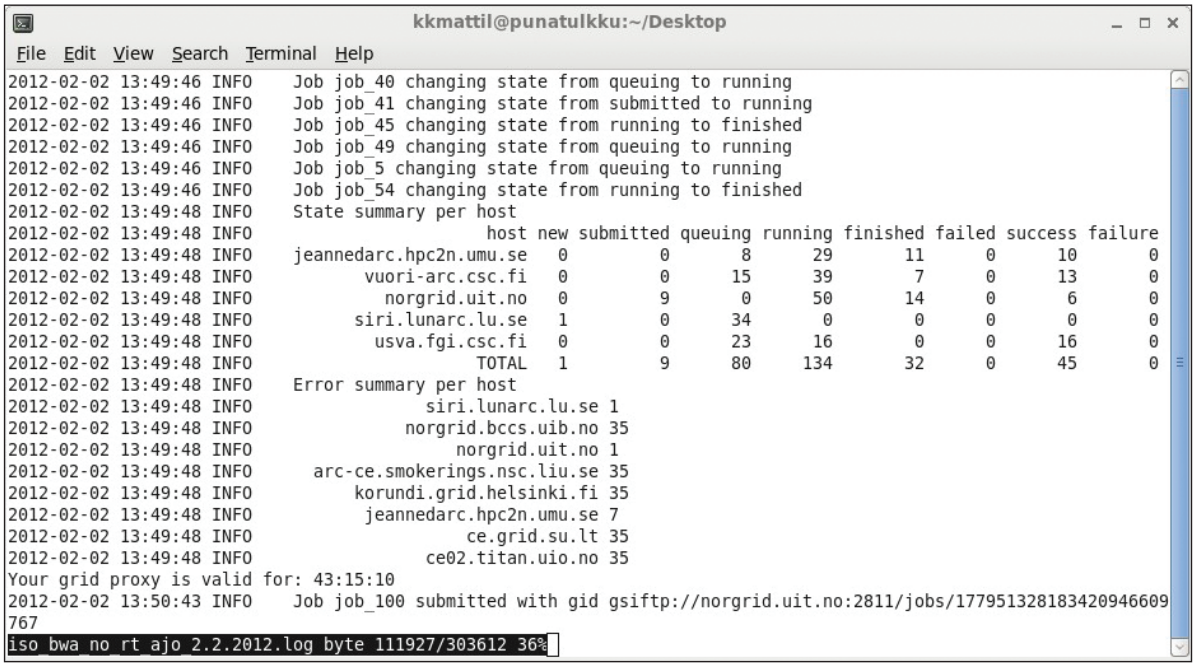


Figure 1. Screen capture of a *grid _ bwa* run. The screen shows the status of a mapping job that is being processed simultaneously in five clusters: two clusters in Finland, two in Sweden and one in Norway. The mapping task was split into 301 sub-tasks. 45 of these sub-tasks are already successfully completed, 32 wait for result retrieval, 134 are currently being executed, 80 are queueing in the clusters, 9 are submitted to the grid and one job waits to be submitted.

```
grid _ bwa aln -query1 reads1.fq -query2
reads2.fq -ref genome.fa -out aln.bam
```

Adding the option *-query2* defines that a paired-end analysis is executed, and that the post-processing is done with the *bwa sampe* command instead of *bwa samse*.

Just like in the previous example, *bwa aln* para-meters can be defined for this command with normal *bwa aln* options. You can also define parameters for the *bwa sampe* command. This is done with options *-sampe _ a* (corresponding to the *bwa sampe* option *-a*), *-sampe _ o*, *-sampe _ n*, *-sampe _ N*, and *-sampe _ r*.

For example, executing a paired-end alignment task, where the seed length is 24 and, in the post-processing state, the maximum insert size is 400 (bwa sampe -a 400), can be defined with command line:

```
grid _ bwa aln -query1 reads1.fq -query2
reads2.fq -ref genome.fa -out aln.bam -l
24 -sampe _ a 400
```

Once the command is launched, it starts printing out log information about the progress of the job (Figure 1). For longer jobs, it is reasonable to forward the *grid _ bwa* output to a separate file and run the command as a background process. This way you can log out and return later on to check how the job has progressed. For example:

```
grid _ bwa aln -query1 reads1.fq -query2
reads2.fq -ref genome.fa -out aln.bam -l
24 -sampe _ a 400 > log.txt &
```

After launching the job, the *grid _ bwa* command must be kept running until all sub-jobs are processed, and the cleaning and post-processing tasks are done.

Accessing the tool

1. Installing the grid _ bwa command

At the moment, this Grid-job submission tool for BWA is in use only in the servers of CSC. So, the easiest way to use these tools is to apply for a CSC user account and join to the Finnish Grid Initiative[1] or the bioscience virtual organisation of Nordic Data Grid Facility (NDGF)[2]. However, grid _ bwa is just small set of tcsh and python scripts. It can be installed on any Linux machine that has the following components: 1) an ARC middleware client; 2) BWA; 3) SAMtools; 4) Python 2.6 or later.

1 http://www.csc.fi/english/research/Computing_services/grid_environments/fgi
2 www.ndgf.org/

Local installation of the EMBOSS (Rice et al., 2000) package is also recommended, as the grid _ bwa uses EMBOSS, if it is available, to check that the input files are in the correct format. The scripts, which form the grid _ bwa tool, can be downloaded from the BWA instruction page of CSC[3]. Further, if you are using other than the FGI Grid environment, you should ask your Grid administrators to install BWA and SAMtools runtime environments onto your ARC clusters. Installation instructions and runtime environment examples can be found form the NDGF runtime environment registry[4].

2. Grid certificates and Virtual Organisations

In addition to the Grid-submission command, the user must have access to some ARC middleware-based Grid environments: e.g., the Grids of NDGF or FGI. These Grid environments, like most middleware-based Grid environments, use personal X.509 certificates for user authentication. Certificates are granted by a Certification Authority (CA), which acts as a trusted third party to ensure that the identity information is valid. For example, Nordic academic Grid users can use the Trans-European Research and Education Networking Association (TERENA)[5] as the certification authority. The certificate is first installed on your Web browser, where it can be used to automatically authenticate you to a Website. You also need to install the certificate on the computing sever that is used to launch the Grid jobs.

Once a researcher has a Grid certificate, he/she can apply for membership of a Virtual Organisation (VO). A VO refers to a group of users or institutions that utilise some Grid resource according to a set of resource-sharing rules and conditions. Typically, VOs focus on some specific branch of science and/or geographic region. A VO is also linked to a distinct set of Grid resources.

At the moment, the tools discussed here can only be used by the members of the Finnish FGI VO however, if needed, the tools can be made available for the NDGF Bio VO, which is open for all Nordic researchers. Researchers working in Finland can join the FGI VO using server[6]. Nordic researchers can join the NDGF Bio VO using server: https://voms.ndgf.org:8443/voms/nordugrid.org/Siblings.do.

Conclusions

The Grid-submission tool described here demonstrates how Grid middleware commands can be embedded in Linux command-line scripts, so that end-users can utilise Grid resources without any knowledge of the Grid middleware. The Grid interface described here performs NGS-read mapping, but similar automatic Grid-submission tools can be set up for other tools too. At CSC, we also provide similar interfaces for SHRiMP, BLAST (Altschul *et al.*,1990), AutoDock (Morris *et al.*, 1998), HHserch (Söding, 2005) and MatLab (MathWorks Inc.).

The Grid interface described here is based on the ARC middleware. In principle, it could also be converted to use other Grid middlewares, but this would require large modifications to the job-managing and data-transport parts of the tool. Further, there already exist BWA implementations that utilise similar approaches for distributing BWA tasks to gLite and BOINC middleware-based grids[7](Luyf *et al.*, 2010).

Acknowledgements

Olli Tourunen is acknowledged for creating the ARC job-manager. This work has been funded by the EGI-Inspire project.

References

1. Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool *Journal of Molecular Biology* **215** (3), 403–410. doi:10.1016/S0022-2836(05)80360-2
2. David M, Dzamba M, Lister D, Ilie L, Brudno M. (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics.* **7**,1011-1012. doi:10.1093/bioinformatics/btr046
3. Ellert M, Grønager M, Konstantinov A, Kónya B, Lindemann J *et al.* (2007) Advanced Resource Connector middleware for lightweight computational Grids. *Future Generation Computer Systems 23*, 219-240. doi:10.1016/j.future.2006.05.008
4. Langmead B, Trapnell C, Pop, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 doi:10.1186/gb-2009-10-3-r25
5. Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760. doi:10.1093/bioinformatics/btp698

3 http://www.csc.fi/english/research/sciences/bioscience/programs/BWA
4 http://gridrer.csc.fi/
5 https://tcs-escience-portal.terena.org/
6 https://voms.fgi.csc.fi:8443/vomses/
7 https://appdb.egi.eu/?#p=L2FwcHMvZGV0YWlscz9pZD02NDE=

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al.* (2009) The Sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079. doi:10.1093/bioinformatics/btp352
7. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Re.*, **18**, 1851-1858. doi:10.1101/gr.078212.108.
8. Luyf AC, van Schaik BD, de Vries M, Baas F, van Kampen AH, Olabarriaga SD (2010)Initial steps towards a production platform for DNA sequence analysis on the grid. *BMC Bioinformatics* **11**, 598-607. doi:10.1186/1471-2105-11-598
9. MathWorks Inc. Natick, Massachusetts, U.S.A
10. Morris, GM, Goodsell DS, Halliday, RS,
11. Huey R, Hart W, Belew RK, Olson AJ (1998), Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function *J. Computational Chemistry*, **19**, 1639-1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B
12. Rice P, Longden I. and Bleasby A. (2000) EMBOSS : The European Molecular Biology Open Software Suite. Trends in Genetics **16**, 276-277. doi:10.1016/S0168-9525(00)02024-2

Gigsaw – physical simulation of next-generation sequencing for education and outreach

David Michael Alan Martin*
College of Life Sciences,University of Dundee, Dundee, United Kingdom

* corresponding author (d.m.a.martin@dundee.ac.uk)

Received 2 May 2012; **Accepted** 21 June 2012; **Published** 15 October 2012
Competing Interests: none

Abstract

Modern sequencing methodologies produce more data in one run than a human being can read in a lifetime. Understanding how such vast quantities of information can be marshalled, assembled and interpreted is a challenging task for students and experienced researchers; it is even more challenging to have to explain this to lay audiences. Abstract representations, such as graphs or algorithms, or practical exercises with 'black-box' software, are limited in cultivating understanding. Gigsaw provides a physical model of next-generation sequencing data that can be readily manipulated, and different algorithms/experiments investigated at the bench top level. It is flexible in application and inexpensive to produce for public-understanding-of-science exercises or undergraduate/postgraduate training.

Availability: **a Web server implementation of the Gigsaw software is freely available at** http://www.combio.dundee.ac.uk/gigsaw/ **and provides the Gigsaw output as PDF aligned for double-sided printing. Source code is available upon request under an open-source license.**

Introduction

Next generation sequencing (NGS) has brought about a paradigm shift in the prosecution of molecular biology research. Modern instruments can produce tens of millions of short DNA reads per day (Metzker, 2010). The conceptual challenge of understanding how to proceed from these tens of millions of sequence reads to a biological interpretation is considerable (Flicek and Birney, 2010). Sequence-assembly algorithms are complex concepts that can be hard for a student to grasp when presented in the traditional form of lectures, and even as practical exercises, where the algorithms are obscured by the quantity of data and 'black-box' software. It can take considerable effort to understand sequence assembly from textbooks or journal articles, an approach that is often beyond many undergraduates, and not suitable for educating the lay public, even though the basic idea of matching sequences is very simple.

Faced with the increasing interest of the public in biological research, and the inclusion of NGS approaches in undergraduate curricula, a new approach was needed to provide an elementary first step in understanding. Two inspirations led to the development of Gigsaw. The first is the 'table of learning' that is attributed to Glasser (Table 1), although the origins remain obscure (Smart and Paulsen, 2011).

Table 1. Quote attributed to William Glasser though the provenance is uncertain.

We learn:
10% of what we read;
20% of what we hear;
30% of what we both see and hear;
50% of what we discussed with others;
80% of what we experienced personally;
95% of what we teach to someone else.

If Glasser's paradigm holds, then providing students with a physical exercise should enable improved understanding over abstract learning by reading or lectures. The second inspiration was an introductory comment by Pevzner and colleagues (Pevzner *et al.*, 2001):

"Children like puzzles, and they usually assemble them by trying all possible pairs of pieces and putting together pieces that match. Biologists assemble genomes in a surprisingly similar way, the major difference being that the number of pieces is larger."

We have developed Gigsaw as a 'Genome jigsaw generator'. It can produce pieces in PDF format that can be readily printed, laminated and used in the classroom, or on the street, for representation of almost any experiment that can be performed with NGS. The physical nature of the model, with the reverse complement appropriately printed on the converse, provides a tangible representation of the abstract concepts behind sequence alignment and assembly. Sequence searching and data-mining become literal concepts that the students can get their hands on. Several example applications are available on the Gigsaw website. Gigsaw has been used in scenarios from bioinformatics conferences with experienced researchers, to public-understanding-of-science (PUS) events with children of all ages who 'get' the concept of building an assembly by matching all the colours very rapidly, often even before they can read.

Implementation

Gigsaw is implemented in Perl as a dual-purpose Common Gateway Interface or command line application. The interface allows almost every aspect of the model to be configured to suit the experiment under consideration. A full list of configurable parameters is given in Table 2. Read length is fixed at 21 bases for single-end simulation. Paired-end simulation has paired reads of

Table 2. Configurable parameters for Gigsaw.

Parameter	Options	Description
Name	Free text	A title for the Gigsaw output
Sequence	1-1,000 characters from the set A,C,T or G	The source sequence from which to derive a Gigsaw.
Number of reads	Positive integer. The output formats 20 reads per A4 page.	All single reads are length 21bp or 10+x+10 for paired end reads
Error rate	Positive integer or 0	The error rate per 1,000 bases. 0 for perfect reads.
SNPs	[0-9]+:[ACTG]+[, [0-9]+:[ACTG]+[, …]]	Specify as position: bases with the number of bases proportional to their prevalence. EG C:T at position 20 in a 3:1 ratio would be 20:CCCT. Separate SNP definitions with commas and/or spaces
Paired-end gap size	0 or positive integer	0 for single-end reads. For paired ends, the actual gap is +- 5%
Sequence colour	X11 or hexadecimal (#FFFFFF) colour	The colour for the read font so multiple experiments can be separated.
Print reference sequence	Boolean	Print a reference sequence ruler from the source sequence.

10 bases each, separated by a distance sampled from a Gaussian distribution with a configurable mean and a standard deviation of 5% of the specified insert size.

Reads are generated as follows: the source sequence is read, and the start point for the read selected at random from a new copy that has been edited according to both the random-error rate selected and any single nucleotide polymorphisms (SNPs) defined. Errors are modelled by a process that randomly samples the genome sequence twice. The first sampling selects, at random, a single nucleotide to replace from the source sequence padded to a length of 1,000 bases, and the second sample randomly selects from the source sequence a replacement nucleotide. This process is repeated until the desired error rate (errors per 1,000 bases) is reached. Not every iteration will induce an error in the source sequence copy, and substitution rates will reflect the base composition of the source sequence. The real error rate is therefore below the requested error rate. Orientations for reads are then assigned randomly.

The desired number of reads (or read pairs) are then printed, 20 to a sheet of A4 paper, with the forward and reverse complement aligned on subsequent pages (Figure 1a and 1b). Upon printing, these can be preferentially laminated for durability, and separated with a guillotine or scissors ready for use.

The source sequence can, if desired, also be generated as a double-sided PDF. This is produced with ruler markings and a one-base pair (bp) overlap at the end of each 21bp segment, allowing the fragments to be joined together into the complete sequence (Figure 2).

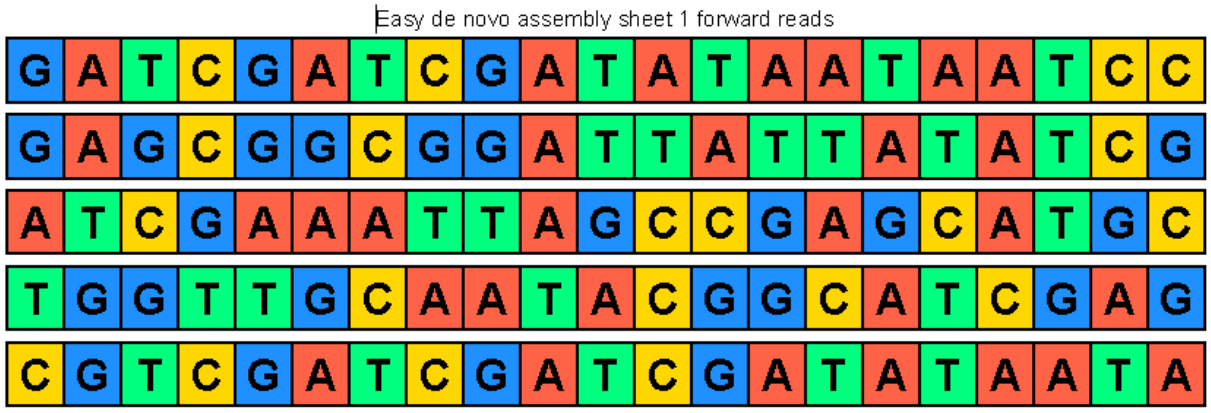


Figure 1a. Panel A: a set of single-end reads. Each time the application is run, a new set of reads is created.

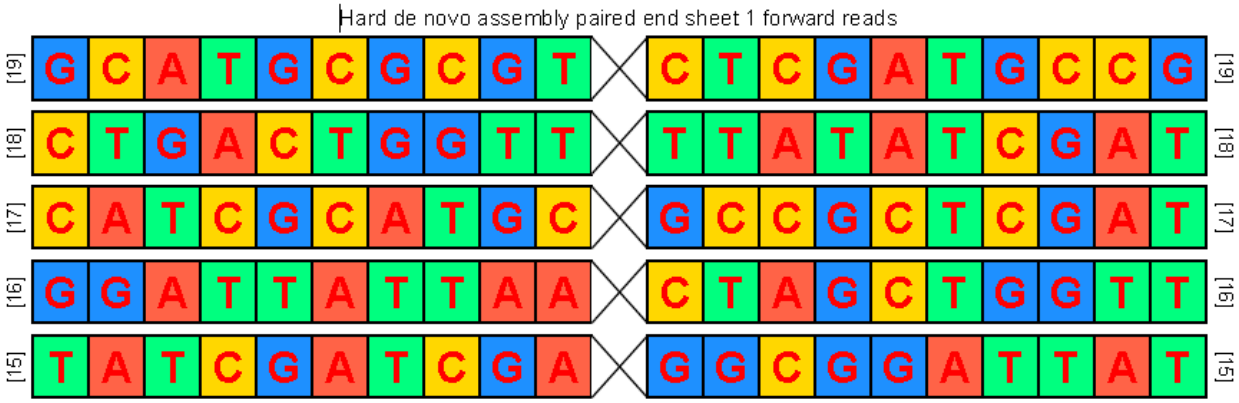


Figure 1b. Panel B: a set of paired-end reads. Each pair is labelled with a unique number. The two central arrows should be pointing towards each other when aligned and be approximately the gap distance apart.

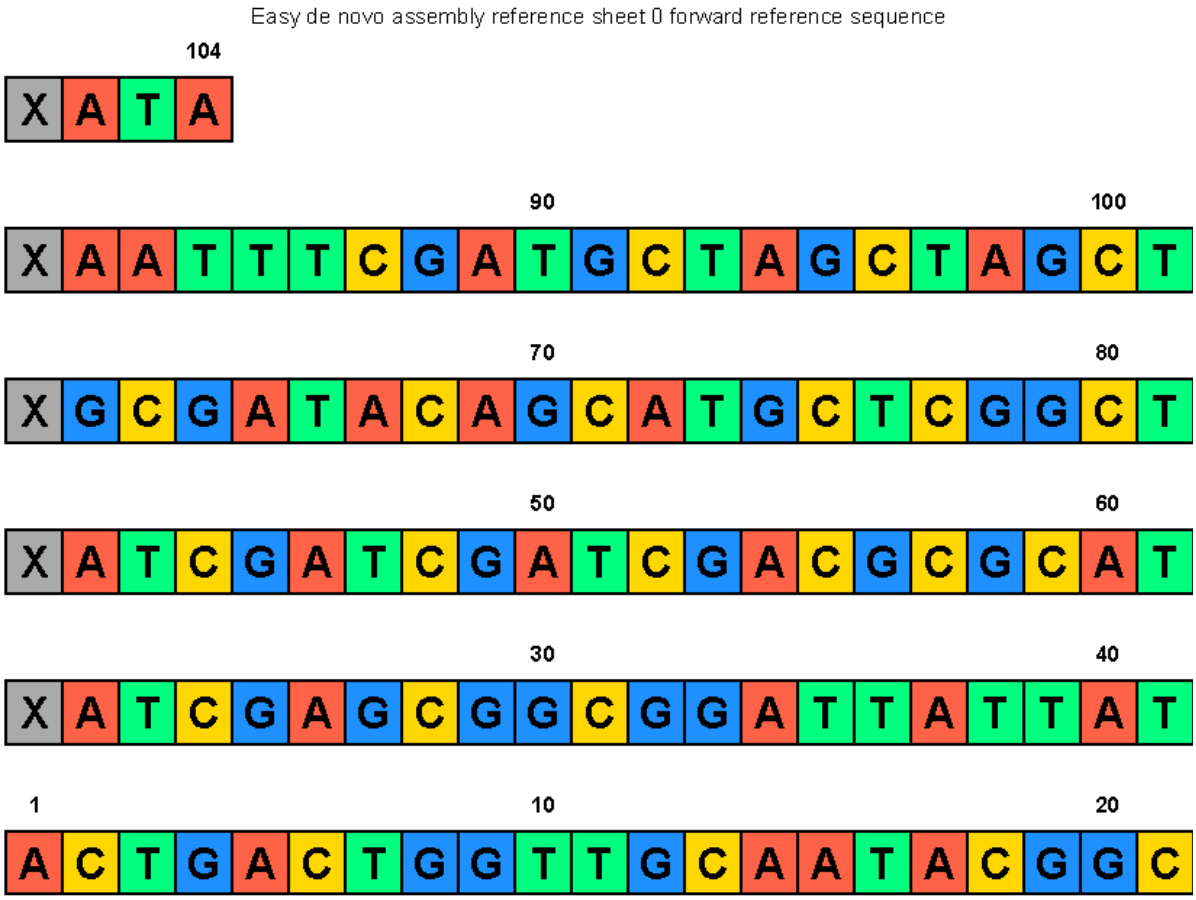


Figure 2. Reference sequence ruler for a short sequence. The grey cross indicates where the previous section should be attached.

Applications

De novo sequencing

Gigsaw is configured to produce reads only with no reference sequence. For PUS exercises, a very low error rate (1% or lower), with a total source sequence length of 50-60 bases, and around 40 reads, works well. For undergraduate exercises, a longer source sequence (up to 150bp) is used, and students are then encouraged to perform database searches with their assembled sequence to try to identify the gene. Care must be taken to avoid repeat regions of longer than about 15 bp, unless the exercise is to illustrate the problem of repeats. Students are asked to consider the evenness of coverage across their sequence assembly as a quality-control measure.

De novo sequencing with repeats

Gigsaw is configured to produce paired-end reads with an insert size that will span the repeats in question. It is also possible to combine single- and paired-end reads by running the Gigsaw application twice, once to generate single-end, and once to generate paired-end reads.

Mutation detection

Gigsaw is run once with the wild-type source sequence to generate a reference genome only. It is then run again with the mutated source sequence to produce reads. Students align the reads to the reference, and identify the mutated residues. A mixture of synonymous and non-synonymous mutations allows mapping onto a disease-related protein of interest, and develops understanding of how sequencing can identify this. Care should be taken to ensure that the students recognise the potential for sequencing errors, so a relatively high error rate (5%) will reinforce the need to see multiple reads with the same mutation. It is probably best to choose smaller proteins, or a single domain with an up-

per bound of about 200bp to maintain interest, while still providing sufficient intellectual stimulation.

Single Nucleotide Polymorphism detection

SNP detection can be performed with a single run of Gigsaw. SNPs are configured by specifying the base position and then the SNP ratio: e.g., for a 3:1 ratio of C to T at position 25, this would be specified as 25:CCCT. Multiple SNPs can be specified. Additional learning points here for the students are the statistical significance of SNP calling depending on the coverage depth and error rate.

RNAseq– intron/exon identification

A reference genome sequence is produced from the source DNA sequence. The RNA sequence is then used to produce a sufficient quantity of reads. It is best to not make the intron too long – anything longer than about 20bp is unnecessary. Students should note that the reads that bridge the splice sites should match both sides.

Quantitative RNAseq

This will require a larger group to get anything reasonably meaningful for analysis. A long genome read approaching the upper limit of Gigsaw (1,000bp) is produced. For each individual 'gene' (probably of 80-100bp) a separate Gigsaw run is required, and the appropriate number of reads generated. These reads are then mixed, and a sample of an appropriate size taken for alignment. Learning points here can include discussion of the detectable dynamic range with respect to read number, and how to deal with multiple matches for a read.

Conclusion

Gigsaw provides an educational tool that is adaptable, durable (if laminated) and extremely cost effective for teaching DNA-sequencing applications. It can be applied in situations from advanced training courses down to public engagement exercises by adjustment of the scale and complexity of the problems presented. Indeed, our experience, though statistically unsound, is that pre-schoolers often perform better than bioinformatics professors at simple *de novo* sequence assembly!

Acknowledgements

The author would like to acknowledge his colleagues and members of the Scottish Next Generation Bioinformatics User Group for helpful suggestions, and Dr. Tom Walsh for expert systems support.

References

1. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* **6**, S6-S12. doi: 10.1038/nmeth.1376
2. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet.* **11**, 31-46. doi: 10.1038/nrg2626
3. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* **98**, 9748-9753. doi: 10.1073/pnas.171285098
4. Smart JC and Paulsen MB (2011) Higher Education: Handbook of Theory and Research. In: Smart, John C.; Paulsen, Michael B. (Eds.). Springer ISBN: 9400707010, Vol. 26, pp. 323.

on sex, drugs and satisfaction

Vivienne Baillie Gerritsen

Pleasure is not a human invention. Experiences that arouse a feeling of contentment are as old as life. They have, in fact, kept life going. It is yet another of Mother Nature's tricks. If an organism perceives something as good, then it will do it again. If you want to keep a species going, the best way to do it is to reproduce. And, if the act of copulation is a pleasant experience, there's a fair chance you'll have another go at it. Eating, sex and social interactions are examples of acts most animals are accustomed to, and for which they are rewarded with a positive feeling. They also happen to be interactions which keep a species alive. But what happens when an animal meets frustration? Following rejection by a potential mate, for example? It finds some other way to quench its desires. Given the chance, *Drosophila melanogaster* will actually turn to alcohol if mating has been denied. Sex and alcohol are part of a highly complex reward system that has had plenty of time to evolve. Recently, scientists discovered the agent which orchestrates both behaviours: Neuropeptide F.

by Thomas Mustaki

Courtesy of the artist

Let there be no misunderstanding: a fly will not spontaneously drown its misery in alcohol when its female counterpart has given it the cold shoulder. But if food soused with alcohol is laid before it, the insect will show a keen preference for it. Because, like sex, alcohol is sensed as a reward. And if a reward is there to take, any animal in its right mind would go for it. Sex is a natural reward; alcohol consumption an artificial one. What the scientists set out to find was whether these two types of pleasure were governed by the same system on the molecular level. It turned out they were. In a nutshell, they discovered that drugs actually hijack the natural reward system. Which explains a lot. It is a discovery that should mark the beginning of important therapies, and which could help people who suffer from afflictions such as stress disorder or drug addiction.

For well over a century now, all sorts of scientists have been trying to understand the fundamentals of animal behaviour. Take Konrad Lorenz and his geese, or Desmond Morris and his naked ape for instance. Why does an animal behave in a certain way? And how? The field is very complex, fascinating and, in some ways, frightening. Is human behaviour, for example, solely determined by the organism's need to survive? Does a child only enjoy an ice-cream because it spells fuel? Or has the pleasure system been diverted somehow? A bit of both no doubt. Tests can now be carried out on the molecular level and behaviourists are able to delve into the parts of animals' central nervous systems that govern given types of behaviour. Thus a neural representation of what is going on is slowly emerging.

Forms of stress – such as sexual rejection or post-traumatic trauma stress syndrome for instance – trigger off certain behaviours that are, more often than not, ruled by a complex reward system. When a male fly plays a love song with its wings, taps its mate gently on the abdomen with its paw, fondly buries its proboscis into the female's private parts and has to suffer rejection, the best way to get over the transient humiliation is to find something that will make it feel

better. On the molecular level, Neuropeptide F (NPF) acts as a sort of 'thermometer of pleasure'. When *Drosophila* is denied copulation, the levels of NPF in its brain are low. When it has been able to mate, the levels are high. Low levels of NPF will make the fly seek out an alternative form of pleasure. A fly with high levels of NPF doesn't feel the need to. Therefore, NPF seems to reflect the state of *Drosophila*'s reward system and a fly's subsequent behaviour.

How did scientists discover the link between copulation, ethanol and NPF? Male flies were isolated with three different types of females: virgins, females that had already mated, and virgins whose heads had been removed (...).The male flies were then offered food that had ethanol in it or not. The flies that had copulated ate both types of food. Those that had suffered rejection turned markedly more to the food soused with alcohol. As did those that had spent time with the headless virgins. Why behead them? This was a way of finding out whether flies suffered from rejection, as opposed to non-copulation. As it turned out, that was not the case. The 'headless virgin' males needed alcohol too. Lack of sex was the answer. Furthermore, frustrated flies that were given a chance to mate, subsequently lost interest in alcohol. So besides the direct link between two behaviours, there is also a mechanism which balances the reward system too, by bringing the levels of NPF back to normal.

The explanation sounds straightforward yet, on the molecular level, the mechanisms are far from understood. NPF and its receptor, yes, are at the heart of the system but how does it work? How does NPF link sexual behaviour with alcohol consumption? NPF is expressed in NPF neurons. The peptide has already been linked to alcohol sensitivity in *Drosophila*, and is known to influence behaviours such as larval intake of noxious foods and physical stress for instance. The novelty here, though, is that a given behaviour actually regulates the levels of NPF. As such, this particular regulation constitutes the basis of *Drosophila*'s reward system. This 'reward system' NPF is probably expressed in different neurons and may be linked to the dopaminergic systems, known to play a major role in reward-driven learning.

How about humans? It is very tempting to draw parallels with the mammalian reward system. Mammals have a similar neuropeptide, known as Neuropeptide Y or NPY, which is involved in the regulation of alcohol consumption. As in *Drosophila*, it is likely that drugs expropriate the human reward system, twisting a natural system into something which becomes harmful to the organism though it is felt as something pleasurable. NPY levels in humans have been shown to be low in the event of depression or post-traumatic stress disorders for instance. In rats, NPY levels have been linked to alcohol consumption and drug taking. But no direct connection has yet been made between social experience, NPY and alcohol consumption. *Drosophila* is not a close relative, yet it undoubtedly offers an excellent model for a greater understanding of the processes underlying drug addiction, alcoholism and obesity to name a few. If NPF is injected into a frustrated *Drosophila*, the insect doesn't feel the need to turn to alcohol any more. Could there be something here for people who suffer from various forms of addiction? Perhaps. But let's not stop eating ice-cream.

Cross-references to UniProt

Neuropeptide F, *Drosophila melanogaster* (Fruit fly) : Q9VET0

References

1. Shohat-Ophir G., Kaun K.R., Azanchi R., Heberlein U.
Sexual deprivation increases ethanol intake in Drosophila
Science 335:1351-1355(2012)
PMID: 22422983

2. Zars T.
She said no, pass me a beer
Science 335:1309-1310(2012)
PMID: 22422968

3. Wen T., Parrish C.A., Xu D., Wu Q., Shen P.
Drosophila neuropeptide F and its receptor, NPFR1, define a signaling pathway that acutely modulates alcohol sensitivity
PNAS 102:2141-2146(2012)
PMID: 15677721

National Nodes

Argentina
IBBM, Facultad de Cs. Exactas, Universidad Nacional de La Plata

Brazil
Lab. Nacional de Computação Científica, Lab. de Bioinformática, Petrópolis, Rio de Janeiro

Chile
Centre for Biochemical Engineering and Biotechnology (CIByB). University of Chile, Santiago

China
Centre of Bioinformatics, Peking University, Beijing

Colombia
Instituto de Biotecnología, Universidad Nacional de Colombia, Edificio Manuel Ancizar, Bogota

Costa Rica
University of Costa Rica (UCR), School of Medicine, Department of Pharmacology and ClinicToxicology, San Jose

Finland
CSC, Espoo

France
ReNaBi, French bioinformatics platforms network

Greece
Biomedical Research Foundation of the Academy of Athens, Athens

Hungary
Agricultural Biotechnology Center, Godollo

Italy
CNR - Institute for Biomedical Technologies, Bioinformatics and Genomic Group, Bari

Mexico
Nodo Nacional de Bioinformática, EMBnet México, Centro de Ciencias Genómicas, UNAM, Cuernavaca, Morelos

Norway
The Norwegian EMBnet Node, The Biotechnology Centre of Oslo

Pakistan
COMSATS Institute of Information Technology, Chak Shahzaad, Islamabad

Poland
Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa

Portugal
Instituto Gulbenkian de Ciencia, Centro Portugues de Bioinformatica, Oeiras

Russia
Biocomputing Group, Belozersky Institute, Moscow

Slovakia
Institute of Molecular Biology, Slovak Academy of Science, Bratislava

South Africa
SANBI, University of the Western Cape, Bellville

Spain
EMBnet/CNB, Centro Nacional de Biotecnología, Madrid

Sri Lanka
Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo, Colombo

Sweden
Uppsala Biomedical Centre, Computing Department, Uppsala

Switzerland
Swiss Institute of Bioinformatics, Lausanne

Specialist- and Assoc. Nodes

CASPUR
Rome, Italy

EBI
EBI Embl Outstation, Hinxton, Cambridge, UK

Nile University
Giza, Egypt

ETI
Amsterdam, The Netherlands

IHCP
Institute of Health and Consumer Protection, Ispra. Italy

ILRI/BECA
International Livestock Research Institute, Nairobi, Kenya

MIPS
Muenchen, Germany

UMBER
Faculty of Life Sciences, The University of Manchester, UK

CPGR
Centre for Proteomic and Genomic Research, Cape Town, South Africa

The New South Wales Systems Biology Initiative
Sydney, Australia

for more information visit our Web site
www.EMBnet.org

EMBnet.journal
ISSN 1023-4144

Dear reader,

If you have any comments or suggestions regarding this journal we would be very glad to hear from you. If you have a tip you feel we can publish then please let us know. Before submitting your contribution read the "Instructions for authors" at http://journal.EMBnet.org/index.php/EMBnetnews/about and send your manuscript and supplementary files using our on-line submission system at http://journal.EMBnet.org/index.php/EMBnetnews/about/submissions#onlineSubmissions.

Past issues are available as PDF files from the Web site:
http://journal.EMBnet.org/index.php/EMBnetnews/issue/archive

Publisher:
EMBnet Stichting p/a
CMBI Radboud University
Nijmegen Medical Centre
6581 GB Nijmegen
The Netherlands

Email: erik.bongcam@slu.se
Tel: +46-18-4716696