An application of peer-to-peer technology to the discovery, use and assessment of bioinformatics programs

To the editor: The emergence of computational biology continues to yield the advancement of diverse algorithmic methodology. Each new advance has typically used unique sources of biological information or applied faster, more accurate heuristics. But identifying the right bioinformatics program for any particular experiment continues to be burdensome as each program's performance is a function of biological context. This is further complicated as most programs can require specialized hardware, large compute resources and disparate software dependencies. It is our hypothesis that a community-driven approach to program discovery, use and assessment will increase the accessibility and relevance of these programs to a wide range of end users. We have developed an open-source, decentralized peer-to-peer platform for bioinformatics analysis, called Chinook, which allows researchers across the globe to freely integrate and access diverse command-line programs, computational resources and databases. We have demonstrated Chinook's utility to the assessment of programs used for transcription factor binding site discovery.

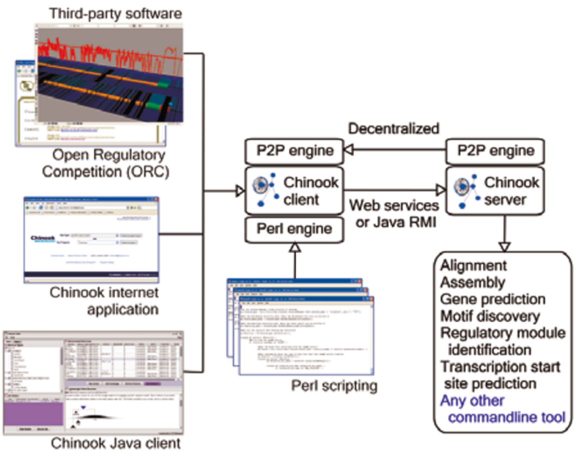

Figure 1 | Chinook platform. Distributed servers advertise command-line algorithms across the internet. Discovered algorithms can be accessed from online, a graphical client, Perl scripting or various Chinook-integrated applications. The ubiquitous availability of algorithmic support through the Chinook Client to various applications reduces the amount of time application developers need to spend (re)integrating support for various algorithms. RMI, remote method invocation.

New programs are integrated and advertised in Chinook using XML or a graphical configuration wizard. End users discover these programs and submit jobs using an internet application, a Java user interface, Perl scripting or through applications using Chinook as an application programming interface (API)[1,2] (**Fig. 1**). The advantages are that program maintenance is supplied by providers instead of end users, redundant advertisement of programs can facilitate distributed computing, and each program can be accessed by a broader spectrum of utilities. To make Chinook amenable to the needs of original program developers[3], each program is advertised with attribution information to increase recognition and citation. Furthermore, Chinook facilitates data entry by integrating the EnsEMBL[4] and Jaspar[5] databases, and has its own plug-in architecture to support the addition of new data sources.

We have applied Chinook to the assessment of computational tools involved in transcription factor binding site discovery using previously established criteria for performance[6]. We have developed a web application called Open Regulatory Competition (ORC), which performs immediate sensitivity, specificity, positive predictive value and combinatorial correlation coefficient measurements for a dynamically discovered set of transcription factor binding site discovery programs across the internet. By using Chinook, ORC is able to quickly analyze newly discovered programs against existing programs for optimal performance against user-defined datasets, thereby uniquely introducing novel selection pressure against a burgeoning population of bioinformatics resources.

There are now over 30 different programs provided through Chinook at locations in Vancouver and in Ottawa Canada, and in Boston, USA; we will continue to add available programs and servers in the future. Chinook is operating-system independent and is licensed under the Lesser GNU public licence (LGPL). Additional documentation and software is available (http://www.bcgsc.bc.ca/chinook).

Stephen B Montgomery, Tony Fu, Jun Guan, Keven Lin & Steven J M Jones

Canada's Michael Smith Genome Sciences Centre, 570 W. 7th Avenue, Vancouver, British Columbia V5Z 4S6, Canada.
e-mail: sjones@bcgsc.ca

1. Montgomery, S.B. *et al. Genome Res.* 14, 956–962 (2004).
2. Wilkinson, M.D. & Links, M. *Brief Bioinform.* **3**, 331–341 (2002).
3. States, D. J. *Nature* **417**, 588 (2002).
4. Birney, E. *et al. Genome Res.* **14**, 925–928 (2004).
5. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. *Nucleic Acids Res.* **32**, D91–D94 (2004).
6. Tompa, M. *et al. Nat. Biotechnol.* **23**, 137–144 (2005).