can extract information critical for new discoveries. Knowledge-based systems will no doubt provide the best opportunity in this regard.

Julie C. Barnes
BioWisdom Limited, Babraham Hall, Babraham, Cambridge, CB2 4AT, UK

Uncovering the complex mysteries of mosaicism

Sir — I read with great interest your News feature "Dual identities", bringing together intriguing results that are starting to emerge in a relatively unrecognized area of human genetics, chimaerism and mosaicism[1]. It did not, however, cover single-gene mutations leading to somatic mosaicism, or germline mosaicism. I would like to draw readers' attention to these developments because the clinical significance of these observations is probably more far-reaching than previously thought.

One of the initial lines of evidence implicating germline mosaicism in single-gene disorders was haplotype-analysis-based detection of the transmission of a muscular dystrophy gene by an unaffected male[2]. Today, we know that a significant number of boys born with Duchenne muscular dystrophy are members of families in which one parent carries the disease-causing mutation only in his or her germ cells, an important issue in genetic counselling. It is increasingly being realized that mosaicism in germ cells can be an important underlying cause of disease in a growing number of genetic disorders[3].

With respect to somatic mosaicism, we are now beginning to understand the biological mechanism by which some boys with X-linked dominant diseases such as Rett syndrome[4] and incontinentia pigmenti[5] survive.

Although it is not clear whether single-gene mutations mainly occur as a postzygotic event or before fertilization at the half-chromatid stage[6], the presence of an X-linked dominant disease in a male is an important diagnostic sign of somatic mosaicism.

Tayfun Özçelik
Ayhan Sahenk Foundation and Department of Molecular Biology and Genetics, Bilkent University, Bilkent, Ankara 06533, Turkey

1. Pearson, H. Nature 417, 10–11 (2002).
2. Darras, B. T. & Francke, U. *Nature* **329,** 556–558 (1987).
3. Zlotogora, J. *Hum. Genet.* **102,** 381–386 (1998).
4. Topçu, M. *et al. Eur. J. Hum. Genet.* **10,** 77–81 (2002).
5. The International Incontinentia Pigmenti Consortium. *Am. J. Hum. Genet.* **69,** 1210–1217 (2001).
6. Gartler, S. M. & Francke, U. *Am. J. Hum. Genet.* **27,** 218–223 (1975).

Bioinformatics code must enforce citation

Sir — Despite repeated calls for the development of open, interoperable databases and software systems in bioinformatics (for example refs 1–3), Lincoln Stein in his Commentary "Creating a bioinformatics nation", with some justification compares the state of bioinformatics to the mediaeval city-states of Italy, and proposes a unifying code of conduct[4]. In considering his proposal, we must ask why such a chaotic situation arose, and why it has been so persistent.

There are many reasons for the existing chaos. Bioinformatics is a rapidly evolving field. Stable interfaces take time to design, implement and maintain. Algorithms and tools evolve and incorporate feedback from users, and the interfaces must necessarily evolve as well. But standards have been developed and widely accepted in other fields undergoing rapid technological change, such as the Internet.

The difference is that academic scientists are responsible for most of the software and data in bioinformatics. Academic careers are advanced by publications that establish priority and citations that validate the impact of the work. Being the first to develop a new approach forms the basis for a peer-reviewed publication, which is not the case for developing and maintaining a standard interface to an old tool or data set. Academic scientists cannot be expected to sacrifice their careers in the interest of community standards. Significant responsibility for standards development and implementation must fall to service organizations such as database providers. These organizations need the support of the academic community in standards development, whereas academic scientists need to benefit from the time and effort they contribute to the process.

Modern bioinformatics software systems are complex. A genome-annotation system, for example, draws on dozens of software components, involving teams of dozens or even hundreds of developers. We need ways to recognize the often-critical contributions of these individuals to the overall result. Stein's code of conduct would facilitate the development of seamlessly interoperable systems in a way that hides the underlying complexity of a calculation from the user. From an academic scientist's perspective, this goal is in direct conflict with the need for recognition and citation and will do nothing for the career of the developer. An academic scientist, therefore, has a strong career imperative to force users to deal directly with their tool or website, and little incentive to make the technology accessible through interoperable systems.

BLAST[5] and FASTA[6] are "citation classics", but they are also at the top of the list of "failure to be cited classics". Projects like NCBI[7] and Ensembl[8] have made useful software tools and large volumes of data widely available, but do not give users the information necessary to cite appropriately the algorithms and software needed to access the system. Ensembl, for example, provides users with alignments performed by BLAST and SSAHA[9] using EST sequences[10,11] aligned to the human genome sequence[12] and gene models created by GenScan[13]. And yet Ensembl lists a citation only to itself on its home page, and the NCBI genome resources pages provide no citation information for the underlying bioinformatics.

If bioinformatics is to emerge as a strong "nation state", Stein's code of conduct needs to address the career imperatives of computational biologists. First and foremost, it must require people to cite their sources. Interfaces and data sets should include explicit citation information, so that systems assembled from components can recursively retrieve citation data from their components and present the user with information for all the modules used in a task.

Graphical user interfaces provide ready mechanisms to display the properties of an object. A user clicking on a gene model should be able to retrieve citation information quickly and automatically for the software and data used to assemble that model. This object- and task-specific citation approach would also provide a mechanism for recognizing the specific contributions of developers in large teams. The use of algorithms, software or data without attribution is plagiarism. Manuscripts that fail to cite bioinformatics sources properly are not acceptable for publication in peer-reviewed journals, and software systems that fail to cite their component sources are not appropriate for use by the scientific community.

David J. States
University of Michigan School of Medicine, Ann Arbor, Michigan 48109, USA

1. Boguski, M. S. *Curr. Opin. Genet. Dev.* **4,** 383–388 (1994).
2. Fasman, K. H. *J. Comput. Biol.* **1,** 165–171 (1994).
3. Karp, P. D. *Trends Biotechnol.* **14,** 273–279 (1996).
4. Stein, L. *Nature* **417,** 119–120 (2002).
5. Altschul, S. F. *et al. J. Mol. Biol.* **215,** 403–410 (1990).
6. Pearson, W. R. & Lipman, D .J. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448 (1988).
7. Wheeler, D. L. *et al. Nucleic Acids Res.* **30,** 13–16 (2002).
8. Hubbard, T. *et al. Nucleic Acids Res.* **30,** 38–41 (2002).
9. Ning, Z., Cox, A. J. & Mullikin, J. C. *Genome Res.* **11,** 1725–1729 (2001).
10. Hillier, L. D. *et al. Genome Res.* **6,** 807–828 (1996).
11. Adams, M. D. *et al. Nature Genet.* **4,** 373–380 (1993).
12. International Human Genome Sequencing Consortium. *Nature* **409,** 860–921 (2001).
13. Burge, C. & Karlin, S. *J. Mol. Biol.* **268,** 78–94 (1997).