Confidence in comparative genomics

Elliott H. Margulies

Genome Res. 2008 18: 199-200
Access the most recent version at doi:10.1101/gr.7228008

Commentary

Confidence in comparative genomics

Elliott H. Margulies[1]

Genome Informatics Section, Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Comparative sequence analysis has become a widespread approach for identifying and characterizing functional elements encoded within genomic sequences. Marked by early successes (for review, see Hardison 2000), a tremendous amount of sequencing capacity has been, and continues to be, utilized for sequencing genomes of related species. Indeed, the choice of genomes selected for sequencing has less to do with the biology or utility of a particular species as an experimental model organism, but rather is guided more by their placement on the evolutionary tree of life. Optimal species are now characterized by an evolutionary distance (typically measured in neutral substitutions per site) that maximizes both sequence alignability and the ability to distinguish neutral DNA from sequences under evolutionary selection. This concept is exemplified in the mammalian species selected for low-redundancy whole-genome shotgun sequencing (Margulies et al. 2005; Green 2007) as well as the 12 fly genomes selected for comparative analyses (*Drosophila* 12 Genomes Consortium 2007; Stark et al. 2007).

With the increased availability of all these species' genomes, various algorithms have been developed to aid in the identification of sequences under purifying selection (Blanchette and Tompa 2002; Boffelli et al. 2003; Margulies et al. 2003; Cooper et al. 2005; Siepel et al. 2005, 2006), which is Nature's way of pointing out sequences that have remained highly similar throughout evolution and have thus been "constrained" for some function, even when we don't know what function that is. Equally interesting are methods to detect genomic sequences under positive selection (Clark et al. 2003; Nielsen et al. 2005; Pollard et al. 2006; Prabhakar et al. 2006; Kim and Pritchard 2007), which highlight rapidly evolving regions that have acquired new functions and might point to functional sequences that make species unique. In addition, comparative sequencing efforts have had a major impact on studies of evolutionary biology, helping to resolve disputed evolutionary relationships and elucidate mechanisms by which evolution has occurred (e.g., Murphy et al. 2001; Nikolaev et al. 2007).

Yet, with all these advances, there still remains a "single point of failure" in the field of comparative genomics—virtually all analyses rely on the generation of a pre-computed multi-sequence alignment. These alignments are typically generated by programs that use a number of computational "short cuts" (such as a progressive alignment approach) to make the task of building genome-wide alignments feasible. While methods that combine the alignment task with other inferences have also been developed (Alexandersson et al. 2003), they are not widely used in large-scale studies because of their complexity and computational cost. Importantly, recent studies have shown that dramatic differences exist between multi-sequence alignments produced by different algorithms (Margulies et al. 2007; Prakash and Tompa 2007), despite the fact that these alignments are attempting to achieve similar goals from the exact same sequence datasets.

The manuscript by Lunter and colleagues in this issue (Lunter et al. 2008) describes an interesting solution to the challenge of imperfect alignments. They first provide a thoughtful and systematic analysis of the three major classes of biases found in sequence alignments: (1) gap/edge wander, resulting from the incorrect placement of gaps due to spurious nonhomologous similarity; (2) gap attraction, resulting in the joining of two closely positioned gaps into one larger gap; and (3) gap annihilation, resulting in the deletion of two indels of equal size for a typically more favorable representation as substitutions. Examples of these three biases are nicely illustrated in Figure 1 of their manuscript (Lunter et al., this issue). From this initial analysis, they conclude that certain regions of pairwise alignments do not have a single theoretically correct solution. Even when the full evolutionary model is known, multiple evolutionarily plausible possibilities exist. Thus, we may never know with certainty the correct homology in certain regions of pairwise alignments.

Their approach to overcoming this challenge is rather elegant and attacks the problem from a different perspective: Instead of trying to get the alignment correct (which they show might not be possible), they "flag" alignment columns that have a high probability of not being correct. While such a solution will not solve the challenges upstream of the alignment process (namely, identifying the correct orthologous sequences to align in the first place), their approach does help negate a major contributor to false-positive/negative results in downstream comparative sequence analyses. It is encouraging that their approach should also be amenable to multi-sequence alignments, since they are typically built up from a series of pairwise alignments.

More than 15% of aligned bases are estimated to be incorrect in currently available whole-genome alignments between human and mouse (Lunter et al. 2008). While modest improvements were made on simulated alignments by more careful modeling of the evolutionary process (in particular, with respect to G + C content and distribution of indel lengths), the majority of alignment errors could not be resolved, reinforcing the need for a probabilistic approach in multi-sequence alignment analyses. These results led them to develop a posterior decoding algorithm that explicitly models uncertainties in inferred alignments. Alignment uncertainty is of particular concern in noncoding regions of mammalian genomes, which are notably difficult to align but also of great interest for identifying regulatory sequences.

With this new "probability of correctness" information that can be assigned to each column of a multi-sequence alignment, one can envision new approaches that incorporate confidence measures in myriad downstream comparative sequence analyses. In essence, we now know which parts of the alignment we can trust and which parts might be suspect—not because the alignment algorithm failed, but because there is no single highly probable result.

[1]Corresponding author.
E-mail Elliott@nhgri.nih.gov; fax (301) 480-3520.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.7228008.

The approach presented by Lunter and colleagues represents an excellent step toward fully probabilistic approaches to alignment and comparative sequence analysis on a genome-wide scale. We can begin to accept the unavoidable uncertainty in multi-sequence alignments and ultimately add confidence into downstream comparative sequence analyses.

Acknowledgments

I thank my colleagues locally and around the world for continued intellectually exciting collaborations. I also thank an anonymous reviewer for helpful comments. This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health

References

Alexandersson, M., Cawley, S., and Pachter, L. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.* **13:** 496–502.

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. 2003. Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* **302:** 1960–1963.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Green, P. 2007. 2× genomes—Does depth matter? *Genome Res.* **17:** 1547–1549.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Kim, S.Y. and Pritchard, J.K. 2007. Adaptive evolution of conserved non-coding elements in mammals. *PLoS Genet.* **3:** e147, doi: 10.1371/journal.pgen.0030147.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., and Hein, J. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Res.* (this issue), doi: 10.1101/gr.6725608.

Margulies, E.H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.

Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17:** 760–774.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294:** 2348–2351.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3:** e170. doi: 10.1371/journal.pbio.0030170.

Nikolaev, S., Montoya-Burgos, J.I., Margulies, E.H., NISC Comparative Sequencing Program, Rougemont, J., Nyffeler, B., and Antonarakis, S.E. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet.* **3:** e2. doi: 10.1371/journal.pgen.0030002.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.

Prabhakar, S., Noonan, J.P., Paabo, S., and Rubin, E.M. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314:** 786.

Prakash, A. and Tompa, M. 2007. Measuring the accuracy of genome-size multiple alignments. *Genome Biol.* **8:** R124. doi: 10.1186/gb-2007-8-6-r124.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Siepel, A., Pollard, K.S., and Haussler, D. 2006. New methods for detecting lineage-specific selection. In *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, pp. 190–205.

Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., Roy, S., Deoras, A.N., et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.