Genome Biology

COMMENT

The DNA60IFX contest

Michael C Schatz[1]*, James Taylor[2,3] and Sven-Eric Schelhorn[4]

We present the full story of *Genome Biology*'s recent DNA60IFX contest, as told by the curators and winner of what turned out to be a memorable and hotly contested bioinformatics challenge. Full solutions, including scripts, are available at http://genomebiology.com/about/update/DNA60_ANSWERS.

The curators' perspective

Michael Schatz and James Taylor

In honor of the 60th anniversary of the publication of the structure of DNA, we organized a contest related to DNA and its applications in current research. The contest began on April 20 and ended on April 25: the anniversary itself, and popularly known as 'DNA Day'. The contest drew nearly 1,000 participants from across the world. Reflecting the transition from genetics to genomics in the 60 years since the discovery, the contest was presented as a series of bioinformatics challenges in which participants would assemble, align or otherwise analyze nucleic acid sequences to identify a message hidden in the data.

The contest consisted of five stages, ordered so that the solution to one stage unlocked access to the next by completing its URL. There were no timing requirements for the first four stages since they were released at a predefined time for all participants, although the overall winners were determined by how quickly they could correctly solve the final stage. The top prize was an iPad, and the second and third place entries had their choice of a one-year subscription to *Genome Biology* or registration to the Beyond The Genome conference. In addition to celebrating the discovery, we hoped to reach out to students and postdocs around the world to motivate them to learn a few new techniques and a few new concepts of molecular biology. This appears to have been quite successful, and several students outside of biology participated in the contest.

The stages of the contest were presented in order of complexity: the first could be completed in a few minutes, while the final stage might require several hours. However all of the stages could be solved by using a combination of open-source software, if only one could identify the correct algorithms to use. The contest problems and solution guide are available online at http://genomebiology.com/about/update/DNA60_INTRO and http://genomebiology.com/about/update/DNA60_ANSWERS respectively.

Stage 1: motif finding

The first stage was based on the common bioinformatics problem of motif finding, such as for identifying a transcription factor binding motif or other regulatory element upstream of a set of gene sequences. Finding true biological motifs requires complex learning approaches such as Gibbs sampling to account for the variability that may be present. For the contest, we simplified the problem to identification of a 7 base-pair sequence motif without any variability or errors. As a result, the solution could be computed in a few seconds with any of a number of *k*-mer counting software packages. Nevertheless, the simplicity of the stage aided in explaining the process of how to use the solution to unlock the subsequent stages, and also made the contest accessible to a very large set of participants.

Solution: TAGCGAC

Recommended algorithm: Jellyfish *k*-mer counter [1]

Stage 2: gene finding

The second stage centered around the important problem of computational gene finding. Users were presented with an artificial one megabase-pair microbial genome, and tasked to identify the open reading frames (ORFs) and analyze their amino acid sequences. ORFs are regions of a genome stretching from a start codon to a stop codon absent of any in-frame internal stop codons, and represent possible protein-coding genes. While not every ORF in a microbe will be a true gene, the longest ORFs typically are, and thus constitute an effective heuristic for training a gene finder for classifying the other ORFs in an unannotated genome [2]. Once the ORFs were identified, participants were then tasked to translate their codons into their corresponding amino acid sequences, and then report the 25th amino acid from the 15 longest ORFs in sorted order. There are

*Correspondence: mschatz@cshl.edu
[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor NY 11724, USA
Full list of author information is available at the end of the article

several gene finding and ORF finding programs available that could be used for solving the stage, including EMBOSS [3] and Glimmer [2], although it seems many participants chose to implement their own given the questions we received, especially to clarify the processing of overlapping ORFs. Care was taken in designing the problem to ensure that the answer was unambiguous by ensuring the top 15 longest ORFs had distinct lengths. Several participants asked if we had a typo in designing the contest, but they should take note that there is no amino acid with the abbreviation 'O'

Solution: THESECRETQFLIFE

Recommended algorithm: EMBOSS getorf program [3]

Stage 3: RNA-seq expression

For stage three, participants were presented with a pair of simulated RNA-seq experiments from a portion of *Escherichia coli*, and asked to find the most highly differentially expressed gene. While RNA-seq has the potential to discover new genes and new isoforms, in this stage we provided the annotation for the genome, and being a microbe, did not include any alternatively spliced genes. As such, identifying the solution was a relatively simple matter of mapping the reads and comparing the mapped read coverage in the two conditions. Curiously, from the access logs it appears at least one person attempted to solve the stage by systematically trying all 93 annotated genes until the correct one was found.

Solution: CARB

Recommended algorithms: Bowtie [4] and SAMtools [5]

Stage 4: 16S metagenomics

Stage four simulated a metagenomics experiment, as used to explore the microbial composition on different sites on the human body or in different environments around the world. A reasonable shotgun metagenomics simulation would have required a larger dataset than was desirable for the stage, thus we chose to simulate a microbial community profiling experiment using amplified 16S rRNA sequences. We randomly selected 80 or so members of the *Helicobacter* genus, together with a matched number from other random genera, from the Greengenes [6] database of 16S sequences. We then generated simulated 250 to 400 base-pair reads from the V1-V3 variable regions, with a progressively decreasing number of reads drawn from each species. Sequencing and other characteristic errors found in real 16S experiments were not simulated after initial evaluations determined they would make the stage difficult to solve unambiguously without a much larger dataset. The resulting dataset was highly enriched for reads from members of *Helicobacter*, allowing an answer to be determined as verified using the RDP classifier [7] or CAMERA [8]. In generating this stage's dataset, we found that if we reduced the prevalence of the dominant genus it quickly became difficult for common taxonomic classifiers to yield an unambiguous answer. However, because *Helicobacter* was so over-represented, the correct answer could easily be guessed just by aligning random reads to an appropriate database.

Solution: HELICOBACTER

Recommended algorithm: CAMERA [8]

Stage 5: decoding the genome

The final stage was to identify a secret message that we had embedded into a genome, and then email us the correct phrase as fast as possible. This simplified the scoring as we had a time-stamped electronic record of the submissions along with the email addresses of the participants. We embedded the secret message using the encoding scheme proposed by Church *et al.*, in which text or images are represented in a binary alphabet expressed in DNA nucleotides [9]. To further complicate the stage, instead of providing the genome with the secret message embedded within it, we simulated the shotgun sequencing of it and presented just the unassembled reads. We expected the participants to then assemble the reads, BLAST the assembly at NCBI to determine the species, align the assembly to the reference, extract the inserted nucleotides, and then decode the message using the included decoder. Alternatively, one could run the decoder script directly on the unassembled reads. The majority of the reads would decode into unintelligible characters, but those with the insertion would decode into recognizable words that could then be assembled into the entire phrase. This approach would be somewhat more complex to implement since most available genome assemblers are specialized for DNA sequences, but has the advantage of skipping the time-consuming steps of assembling and BLASTing to determine the reference. Indeed, the winning entry used this shortcut to outpace the competition.

Solution: 'We went up, saw the structure, we came back to King's and looked at our Pattersons, and every section of our Pattersons we looked at screamed at you, "Double Helix!" And it was just there! - once you knew what to look for. It was amazing.' (a quote from *Genome Biology*'s DNA Day interview with Ray Gosling [10])

Recommended algorithms: ALLPATHS-LG [11], BLAST [12] and MUMmer [13]

The first correct solution to the final stage was emailed just 19 minutes after posting the challenge by Sven-Eric Schelhorn of the Max-Planck-Institut für Informatik, Germany. The second place winner was physics undergraduate Kevin Wang at the University of Chicago, USA just seconds behind, and the third place was Gustavo Lacerda at the Campinas State University, Brazil in 24 minutes. Twenty-four participants emailed the correct solution to the final challenge in the three hours before we announced the winner, hundreds made it to that stage, and nearly 1,000 participants completed at least the first stage of the contest. Participants came from across the entire globe and most were at academic or research institutions.

Interestingly, the number of bioinformatics competitions is on the rise, including the DREAM [14], Assemblathon [15,16] and Sequence Squeeze [17] contests to name but a few [18]. This rise reflects the increased availability of datasets, the increasing diversity of problems and approaches in the field, and perhaps even the competitive nature of bioinformaticians to strive for the best method to solve a given problem. A well-designed contest provides a unique mechanism for broad evaluation on a level playing field, especially when a well-defined gold standard is available. To that end, here we posed artificial problems with specific correct answers, although the problems had the same form as might be seen with genuine data.

We have organized similar contests to DNA60IFX for the last several years at Beyond The Genome, but this was our first all-electronic contest. Details of the contest were broadcast using Twitter, email and blogs, although it appears most participants learned of the contest over Twitter. Without a physical presence, the contest partially lacked the sense of head-to-head competition that we had seen at Beyond The Genome, but we were able to reach a much broader audience than ever before. Overall, we feel this was a worthwhile trade-off, and enabled us to more directly reach our target audience. In addition, Twitter was extremely useful for rapid impromptu discussion between the participants and for clarification of the rules. Given the success of the project, we are already planning the next contests for later this fall and are also considering making the DNA Day challenge an annual event.

See you at Beyond the Genome (http://www.beyond-the-genome.com/) on October 1-3 for the next contest!

The winner's perspective

Sven-Eric Schelhorn

I really enjoyed taking part in the DNA60IFX challenge. The curators and *Genome Biology* offered a well-planned competition (barring technical glitches) and several elegant problems for the participants to solve. Public events such as these are a great way to increase visibility and uptake of bioinformatics. Apart from the global benefits of the competition, I also took great personal enjoyment in hacking (and, in the last case, breaking) the problems. Solving these kinds of challenges (although perhaps in the more serious and difficult setting of actual science) is why we computational biologists are in the game, after all. Doing it under time pressure was absolutely exhilarating.

My success in this competition is based to a large extent on luck – in particular the fact that part of my PhD thesis (which I plan to hand in soon, I promise…) deals with the detection of insertions of potentially unknown viruses in whole human cancer genomes. Luckily for me, this thesis work resulted in some nice computational tools that, in hindsight, proved to be well-suited to solving the last stage of the DNA60IFX challenge. So that's how I solved the last problem (and there are better ways, see below): after downloading the raw data, I assembled the sequencing reads *de novo* (using the beautiful Ray assembler, published in *Genome Biology* [19]). I knew that, with any luck, the resulting scaffold would contain the insert that coded for the secret message. While running the assembly, I also applied a pretty fast taxonomic annotation pipeline that I had developed as part of my thesis, in order to determine if the reads originated from any organism that had already been sequenced (for the nerds: by mapping against a combined index containing microbial and fungal Refseq genomes, as well as all known 16S ribosomal RNAs).

By about two minutes into the challenge, I had both a draft *de novo* assembly of the sequencing reads and a taxonomic classification (also based on the raw reads) pointing to *Wolbachia*. Quite an interesting parasite, by the way. Now, as mentioned before, finding a long insert in a *de novo* assembly can be very difficult. In general, one has only two options: (1) use the organism's known reference genome and compare it to the assembly using whole-genome alignment (optionally, this can also be done without an assembly by mapping the sequencing reads against the reference, finding the insert breakpoints by paired-end analysis, and then micro-assembling the insert's content based on unmapped reads); or, (2), use information about the insert itself in order to identify it directly in your *de novo* assembly or within the original reads.

Given that the solution to the challenge did not require identifying the *Wolbachia* breakpoints (the location of the insert in the genome), I opted for the second strategy since it would potentially tell me the sequence of the insert in less time compared to the reference-based approach. Luckily, the DNA60IFX insert had a clear structure: it coded for a text message. So I spent the next ten minutes adapting an old script of mine to take the *de novo* assembly and translate all its sub-sequences of a certain length (*k*-mers) into text representations. I used *k*-mers to limit the effect of possible frame shifts, which in the end proved to be unnecessary due to the high quality of the sequencing data. Writing the script was somewhat complicated by my daughter, who, perhaps afflicted by the excited state of her father, was trying to support me by randomly punching the keyboard and trying to climb on me (my spouse was busy at the time and I was 'in charge'). Handing over my mobile phone to the offspring pacified her for some precious minutes.

At first, most of the sequence generated by the script was gibberish, so I instructed it to only keep fragments that yielded printable characters (letters and numbers). After some random fiddling with the *k*-mer parameter (that is, 'empirical optimization'), I retrieved a long text message that looked legit from the output – eureka!

After reading it, I had to smile from ear to ear: while illustrating a beautiful moment of scientific history, the secret message also displayed two additional facets of the discovery of the DNA structure that are often overlooked. One is a tribute to Rosalind Franklin, in whose lab the quote originated (or so I assumed), and who received less than her rightful share of the merits of this discovery. The second facet is that the quote itself is attributed to Raymond Gosling, Franklin's PhD student. Ah, I mused, there we have it – the person who did all the actual work, late at night in the lab, and consequently was first to see the decisive evidence. The lowly PhD student, one of us! Under the impression that these were twenty minutes productively spent, I emailed in the solution.

Only after two more minutes did I realize that the same script would also have worked on the raw input data (the reads) since the *k*-mer approach makes it applicable to the short read length of the dataset. Indeed, I was able to retrieve the same text (although in smaller fragments) directly from the raw data without any assembly or reference – just by computational brute force – within less than a minute. Oh well, I thought, stupid me spending all the precious time on assembly and taxonomic annotation, surely all the smart people participating in the competition were taking the direct approach all along and I would never win. And so, when my spouse returned home half an hour later, I told her that I was too slow to have a chance at winning, but that it had been enjoyable all the same. Well, I was partly right.

Acknowledgements

MCS and JT: We would like to thank Clare Garvey, Naomi Attar, and all at *Genome Biology* for sponsoring the contest and helping us organize it. We would also like to thank David Dooling and Ben Langmead for helping us organize the bioinformatics challenges at the past Beyond The Genome conferences, and Steven Salzberg for inspiring the sequence challenges used this year. Finally, we would like to thank all of the participants for your effort completing the contest and pointing out any problems with the directions.

SES: I would like to thank T. Lengauer and the Max-Planck Society for providing computational resources used to solve the challenge.

Author details

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor NY 11724, USA. [2]Department of Biology, Emory University, Atlanta GA 30322, USA. [3]Department of Mathematics and Computer Science, Emory University, Atlanta GA 30322, USA. [4]Computational Biology and Applied Algorithmics Department, Max-Planck-Institut für Informatik, Campus E1 4, 66123 Saarbrücken, Germany.

Published: 28 June 2013

References

1. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers**. *Bioinformatics* 2011, **27**:764–770.
2. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer**. *Bioinformatics* 2007, **23**:673–679.
3. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**:276–277.
4. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**:R25.
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.
6. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Appl Environ Microbiol* 2006, **72**:5069–5072.
7. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol* 2007, **73**:5261–5267.
8. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J: **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource**. *Nucleic Acids Res* 2011, **39**(Database issue)**:**D546–551.
9. Church GM, Gao Y, Kosuri S: **Next-generation digital information storage in DNA**. *Science* 2012, **337**:1628.
10. Attar N: **Raymond Gosling: the man who crystallized genes**. *Genome Biol* 2013, **14**:402.
11. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data**. *PNAS* 2010, **108**:1513–1518.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–410.
13. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biol* 2004, **5**:R12.
14. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G: **Wisdom of crowds for robust gene network inference**. *Nat Methods* 2012, **9**:796–804.
15. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung W-K, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol İ, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, *et al.*: **Assemblathon 1: a competitive assessment of *de novo* short read assembly methods**. *Genome Res* 2011, **21**:2224–2241.
16. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol İ, Boisvert10 S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis É, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, *et al.*: **Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species.** *arXiv e-print* 2013, arXiv:1301.5406v2.
17. Holland RC, Lynch N: **Sequence squeeze: an open contest for sequence compression**. *GigaScience* 2013, **2**:5.
18. **Overtaken by events**. *Nature* 2013, **497**:535.
19. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J: **Ray Meta: scalable *de novo* metagenome assembly and profiling**. *Genome Biol* 2012, **13**:R122.

doi:10.1186/gb-2013-14-6-124
Cite this article as: Schatz MC, *et al.*: **The DNA60IFX contest.** *Genome Biology* 2013, **14**:124.