The Metastructure Model

Zachary King Ali Ebrahim

August 19, 2015

What is a metastructure model?

A metastructure model is:

- A genomic knowledge base
- A collection of genome-oriented annotations that include:
  - Genes
  - Transcription units (TU)
  - Transcription start sites (TSS)
  - Ribosome binding sites (RBS)
  - (Translation pause sites)
  - Sigma factor binding sites
  - Transcription factor binding sites
- A self-consistent representation of the genome with rules for consistency

The Metastructure model includes the content of a traditional genome annotation, but it has a different purpose. Eventually, the Metastructure model should include all the knowledge about a genome that has positional information.

Experimental error is considered in the model by assigning a confidence interval for each locus in the model.

A Metastructure model is constructed using a reference genome annotation and omics datasets. A workflow can be described for this process. It includes a number of manual and automated steps.

Rules for consistency are defined. These rules identify groups of features whose locations are not consistent with our general understanding of the genome. The inconsistencies identified by these rules require manual curation. They lead to improvements in the model and improvements in our understanding of genome structure and organization.

From genome annotation to metastructure

Genome annotations were the first product of the sequencing era. Early annotations, such as the genome annotation for *E. coli* (Blattner et al. 1997), included genes, operons, regulatory sites, mobile genetic elements, and repetitive sequences in the genome.

Originally, features were identified at the genome scale using "sensors" that scanned for hallmark signals of a feature like a transcription start site or using sequence homology with genes that had already been annotated by more tedious means (Stein 2001).

The SEED project promotes an approach to genome annotation that is based on *subsytems* (R. Overbeek et al. 2005).

The NCBI RefSeq database houses genome annotations, and NCBI has a genome annotation pipeline to generate annotations for new genome sequences (Pruitt et al. 2012).

To build upon genome annotations—which have been referred to as *one dimensional*—it is possible to build genome-scale models (GEMs) of metabolic pathways and their constituent biochemical reactions, reactants, and genes. GEMs are a second dimension of annotation (Reed et al. 2006). Beyond just annotation, GEMs provide a mathematical structure and can actually predict biological phenotypes (Bordbar et al. 2014). Therefore, they can be improved by comparing model predictions to experimental observations and then systematically addressing the failed predictions (Bordbar et al. 2014; Reed et al. 2006).

(Reed et al. (2006) also brings up the spatial localization of genomes, calling it 3D annotation or ultrastructure. Is this part of the metastructure model? Figure 1 in that paper is pretty strange.)

Mathematical models of the genome do not yet exist. We propose a mathematically structured model of the genome: a *metastructure model*.

Notes

- the term "metastructure" has been used for proteins [10.1007/s00018-009-0117-0], but it was also discussed with the current meaning in SB1 [1107038855]
- other references: Aziz et al. (2008), Cho et al. (2009), Mendoza-Vargas et al. (2009), Cho et al. (2012), Harrow et al. (2012), Salgado et al. (2013), Karolchik et al. (2014)

References

Aziz, Ramy K, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, et al. 2008. "The RAST Server: rapid annotations using subsystems technology." *BMC Genomics* 9: 75. doi:10.1186/1471-2164-9-75.

Blattner, F, I Plunkett G, C Bloch, N Perna, V Burland, M Riley, J Collado-Vides, et al. 1997. "The Complete Genome Sequence of *Escherichia coli* K-12." *Science (80-.).* 2771613 (September): 1453–62.

Bordbar, Aarash, Jonathan M Monk, Zachary A King, and Bernhard Ø. Palsson. 2014. "Constraint-based models predict metabolic and associated cellular functions." *Nat. Rev. Genet.* 15 (2): 107–20. doi:10.1038/nrg3643.

Cho, Byung-Kwan, Stephen Federowicz, Young-Seoub Park, Karsten Zengler, and Bernhard Ø Palsson. 2012. "Deciphering the transcriptional regulatory logic of amino acid metabolism." *Nat. Chem. Biol.* 8 (1): 65–71. doi:10.1038/nchembio.710.

Cho, Byung-Kwan, Karsten Zengler, Yu Qiu, Young Seoub Park, Eric M Knight, Christian L Barrett, Yuan Gao, and Bernhard Ø Palsson. 2009. "The transcription unit architecture of the *Escherichia coli* genome." *Nat. Biotechnol.* 27 (11). Nature Publishing Group: 1043–9. doi:10.1038/nbt.1582.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. "GENCODE: The reference human genome annotation for the ENCODE project." *Genome Res.* 22 (9): 1760–74. doi:10.1101/gr.135350.111.

Karolchik, Donna, Galt P. Barber, Jonathan Casper, Hiram Clawson, Melissa S. Cline, Mark Diekhans, Timothy R. Dreszer, et al. 2014. "The UCSC Genome Browser database: 2014 update." *Nucleic Acids Res.* 42 (D1): 764–70. doi:10.1093/nar/gkt1168.

Mendoza-Vargas, Alfredo, Leticia Olvera, Maricela Olvera, Ricardo Grande, Leticia Vega-Alvarado, Blanca

Taboada, Verónica Jimenez-Jacinto, et al. 2009. "Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*." *PLoS One* 4 (10). doi:10.1371/journal.pone.0007526.

Overbeek, Ross, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, et al. 2005. "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." *Nucleic Acids Res.* 33 (17): 5691–5702. doi:10.1093/nar/gki866.

Pruitt, Kim D., Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. 2012. "NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy." *Nucleic Acids Res.* 40: D130–35. doi:10.1093/nar/gkr1079.

Reed, Jennifer L, Iman Famili, Ines Thiele, and Bernhard O Palsson. 2006. "Towards multidimensional genome annotation." *Nat. Rev. Genet.* 7 (2): 130–41. doi:10.1038/nrg1769.

Salgado, Heladia, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S. García-Sotelo, Verena Weiss, et al. 2013. "RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more." *Nucleic Acids Res.* 41 (D1): 203–13. doi:10.1093/nar/gks1201.

Stein, L. 2001. "Genome annotation: from sequence to biology." *Nat. Rev. Genet.* 2 (7): 493–503. doi:10.1038/35080529.