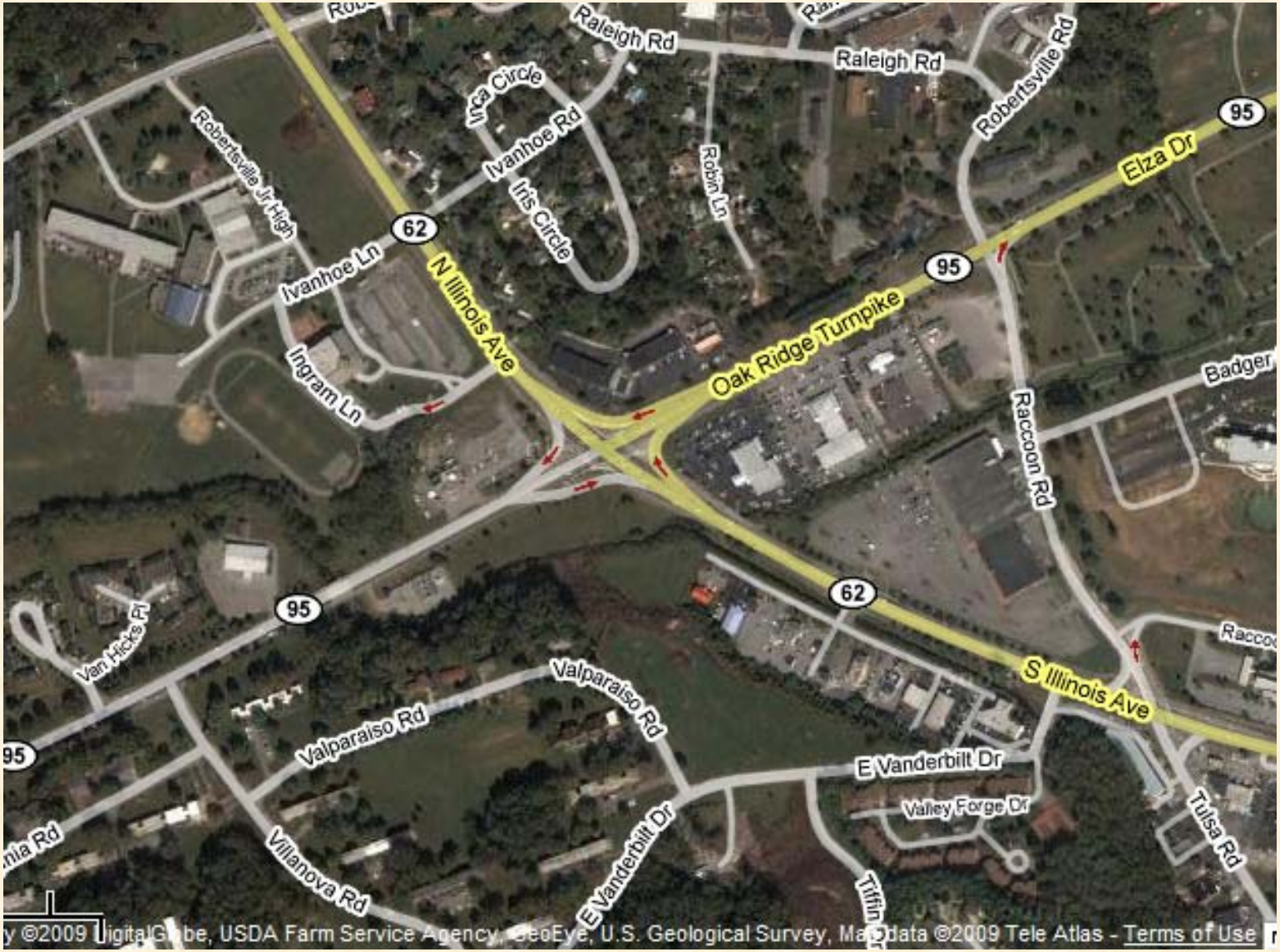
Automated Microbial Genome Annotation: the current state and future challenges

Miriam Land

Sequencing, Finishing and Analysis in the Future Meeting
May 27-29, 2009

How and Why to Annotate?

Ask the locals

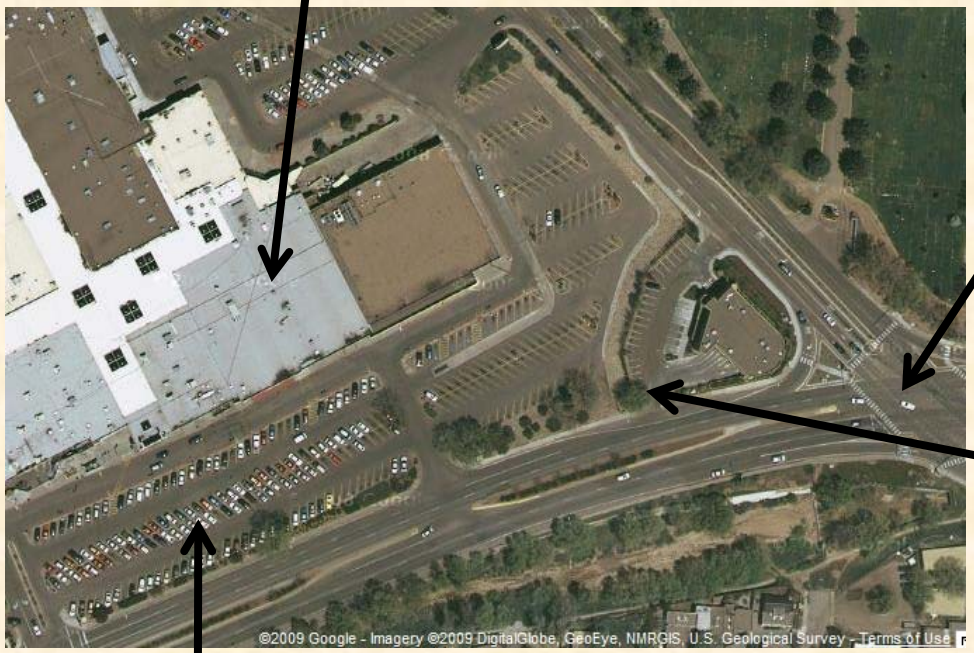


Human recreation center

Sequence Similarity Searches

- **BLAST against your favorite database**
  - **NR**
  - **UniProt**
  - **KEGG**
- **Pros**
  - **Large databases for comparison**
- **Cons**
  - **All amino acids equally important**
  - **Many different contributors, many points of view, many differences in experience**

Covered retreat



Major transportation hub

Season-dependent greenery

Mobility elements

- **Shopping mall???**
- **Major employer???**
- **High School???**
- **Hotel???**

Domain-type searches

- **Interpro**
  - **Pfam, TIGRfam, Smart, etc.**
- **RPS-BLAST**
  - **COGs – Groups of proteins**
  - **PRIAM – Enzyme categories**
- **Pros**
  - **Different weights to different amino acids**
  - **Smaller list of curators with succinct descriptions**
- **Cons**
  - **Smaller databases**

Other Genomic Context Clues

- **Metabolic pathway test**
- **Taxonomy test**
- **Phylogenetic test – someday**
- **Protein folding tests – someday**
- **Experiments – not integrated**

Current State of Genome Annotation

- **Lots and lots of genomes every year**
- **Push to increasing automation**
  - **faster and more likely to be consistent**
  - **computers don't play favorites**
  - **always interprets the same information the same way**
- **The "field guides" vary in methodology and applicability**
- **What goes around comes around**
- **The end result is a hypothesis**

What is the default annotation?

Maps

- **Roads**
- **Parks**
- **Major attractions**

Genomes

- **Protein coding genes (Prodigal)**
- **tRNAs (tRNAscan-SE)**
- **rRNAs (Rnammer)**
- **A label (product) for each gene – field guides**

Not everyone interested in the same annotation

Other Annotation Interests

- **Repeats**
  - **simple direct repeats**
  - **dispersed duplicated segments**
  - **CRISPRs**
  - **transposons/insertion elements**
- **Miscellaneous RNAs**
- **Promoters**
- **Terminators**
- **Other regulatory elements**

Quality depends on individual components and the relationships

Annotation Pipeline Output

- **Web site**
  - **global overview**
  - **drill down to the genes**
  - **tab-delimited formatted files**
- **GenBank files**
- **Upload into IMG**
- **Existing Tools**
  - **Regulator tool**
  - **Transporter tool**

Future of Annotation Pipeline

- **Standards for annotation**
  - **"Standards in annotation next big challenge and deemed impossible" – Patrick Chain**
- **Quality assessment**
  - **Currently all data is presented as uniform in quality**
- **Standardize vocabulary**
- **More tools for consistency and speed**
- **Web-based annotation tool**

"Annotation Quality Sequence" deserves annotation of comparable quality

- **"1/3 of annotation is good quality, 1/3 is marginal, and 1/3 is unknown" – Dan Drell**
- **Currently our best measure of quality is a consensus of the "field guides"**
- **Currently do better with the universal**
- **Standards will be the key to knowing the difference**

Credits

- Loren Hauser,
- Frank Larimer,
- Yun-Juan (Janet) Chang,
- Cynthia Jefferies,
- Gwo-Liang Chen, and
- Bob Cottingham
- Folks at IMG and GenBank