To Boldly GO…

Amelia Ireland
GO Curator
EBI, Hinxton, UK

A Brief History Of GO

Past:

- Began in 1998 as a collaboration between FlyBase, the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD)
- About 3800 terms by 1999
- Ontology text files edited by hand (!)

A Brief History Of GO

Present:

- GO Consortium includes 20+ genome databases
- Used by many groups in academia and industry
- Nearly 18000 terms
- Four full time GO curators
- Many tools and software
- GO paradigm much imitated

OBO

- Web-based repository for open biological ontologies
- Five criteria:
  - Open; no licensing or fees
  - Use common shared syntax
  - Orthogonal to existing OBO ontologies
  - Unique identifiers / namespace
  - Definitions for terms

OBO

http://obo.sf.net/

Cross Products

- Use GO in combination with other vocabularies to create more complex concepts

Extension and Integration of the Gene Ontology: Combining GO vocabularies with external vocabularies.

Hill DP, Blake JA, Richardson JE, Ringwald M. 2002.

Genome Res **12:** 1982–1991

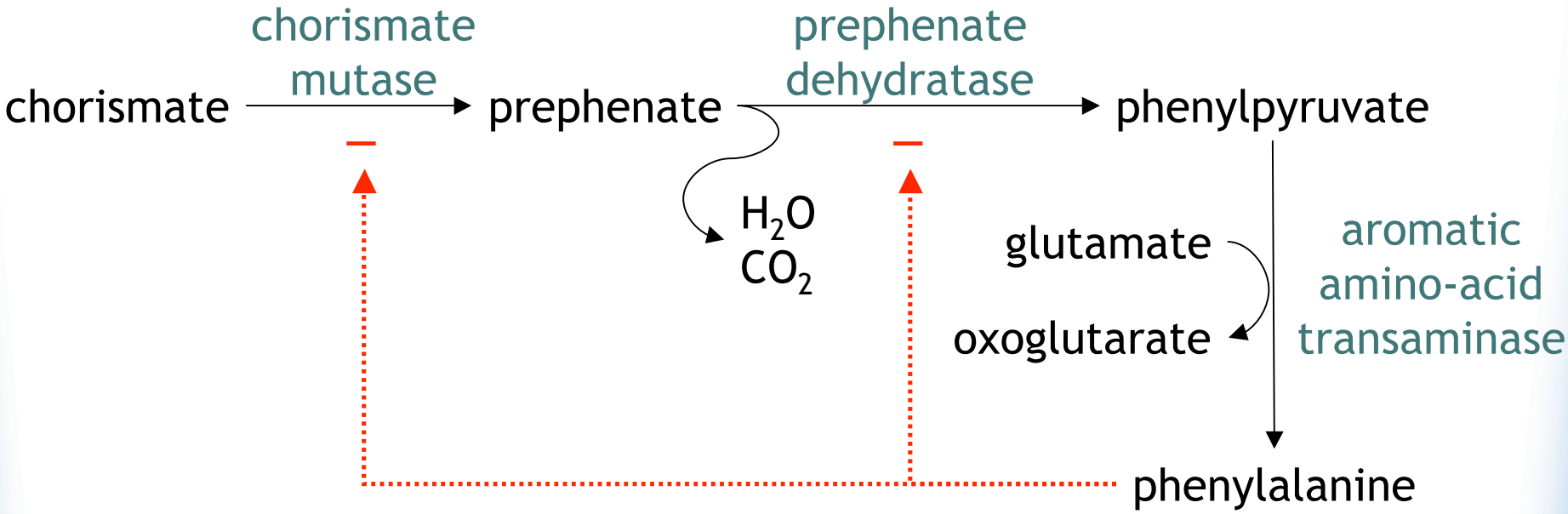
Cross Products

- GO has three ontologies
  - Biological process
  - Molecular function
  - Cellular component
- Extend by combining with terms from other vocabularies

Cross Products

- Narrative method: create terms manually as needed

  phenylalanine biosynthesis



Cross Products

- Combinatorial approach: create all combinations of terms (preferably using a script!)

  phenylalanine biosynthesis

  - biological process ontology
    - metabolism, biosynthesis, catabolism, regulation
  - biochemical ontology
    - chemicals involved in pathway

Cross Products

Demo

Cross Products

- Combinatorial method more thorough but may produce unwanted terms
- Can also lead to massive term proliferation
- Quality of terms (and definitions) depends on source ontologies
- May be better to create cross products as a separate ontology or during annotation

Term Decomposition

- Parsing of GO terms
- Work in progress; Chris Mungall, BDGP
  http://www.fruitfly.org/~cjm/obol-0.02/doc/obol-doc.html

Term Decomposition

- Many GO term names have a regular structure:

  [compound] binding

  [anatomical part] morphogenesis

  regulation of [process]

  x biosynthesis from y

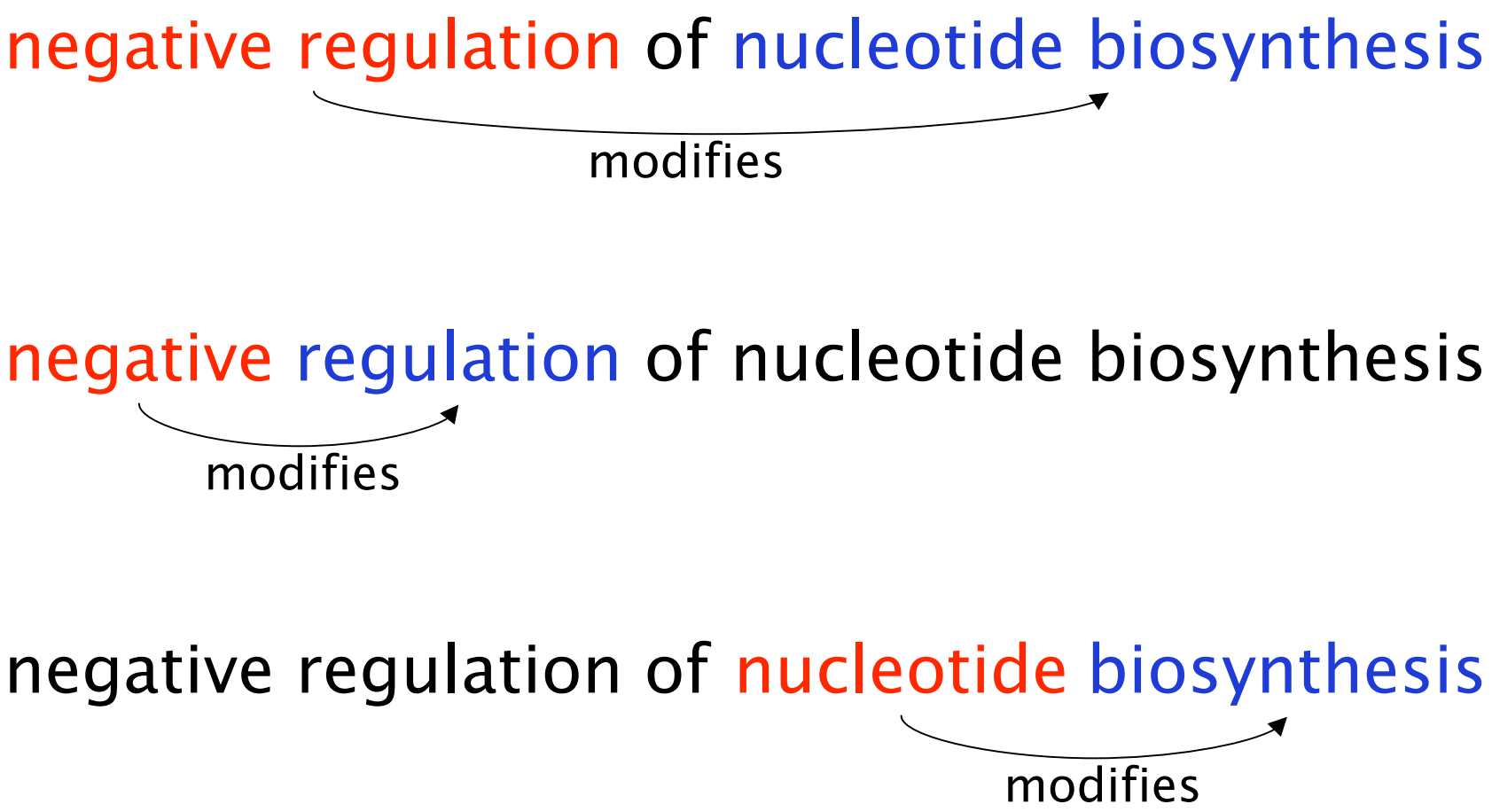
  x biosynthesis, z pathway

- These GO term strings follow consistent implicit naming rules

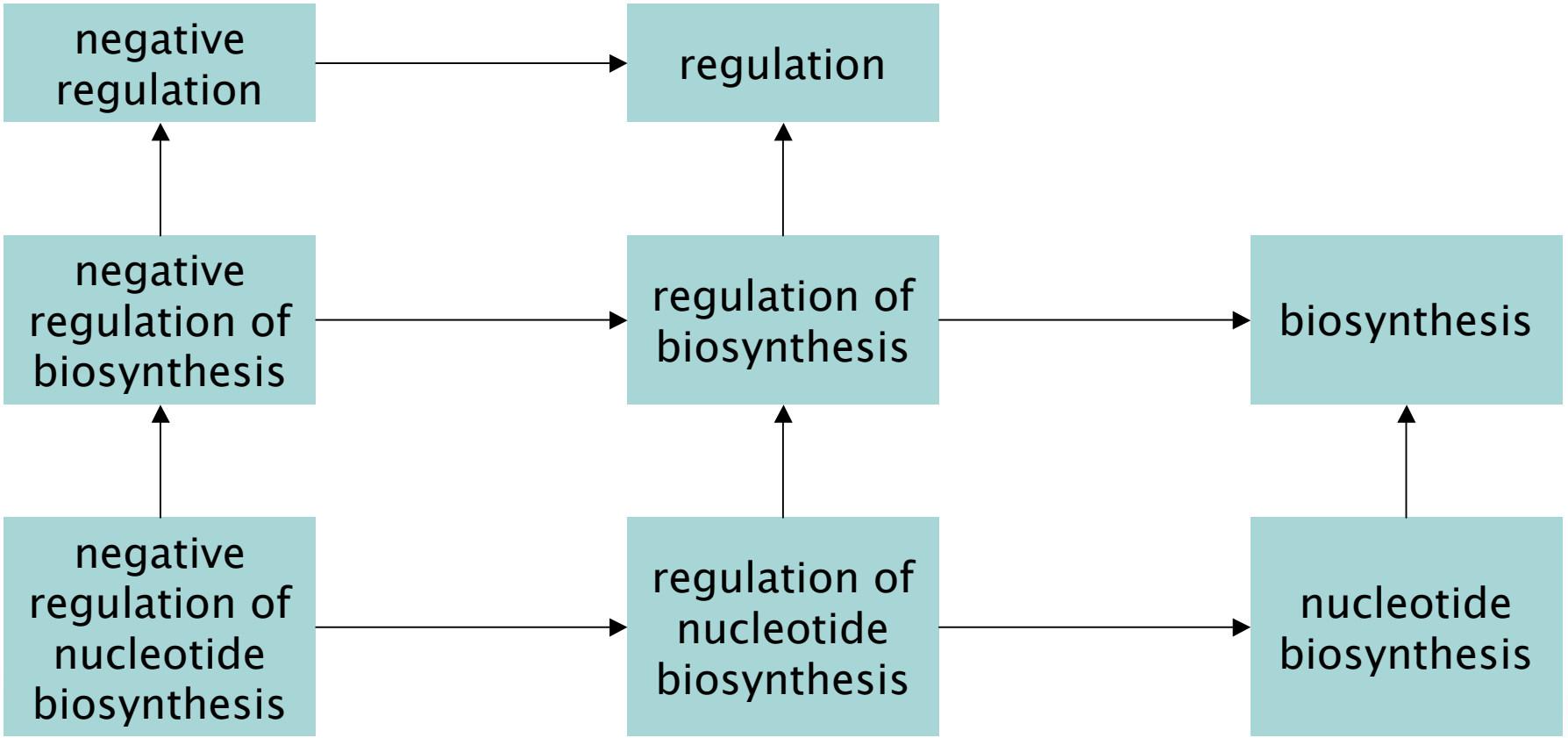
Term Decomposition

- Formal grammar: a rule system for parsing (decomposing) and generating (composing) sequences of symbols
- Using an English language grammar, should be able to parse GO term strings into **tokens** and generate new GO term strings from these tokens
- **Definite Clause Grammar** used as it can be augmented with additional logical constraints; implemented in Prolog

Term Decomposition



Term Decomposition



Term Decomposition

- Over **40%** of GO terms can be (at least partially) decomposed
- These can then be linked to terms from other OBO ontologies – anatomy, biochemistry, cell type, etc.
- Missing GO terms and relationships suggested
- Can also be used to suggest terms in other OBO ontologies

Term Decomposition

- Some standardization required
  - *cytosol* vs *cytosolic*
- Terms with multiple parses require biological knowledge
  - **smooth** muscle contraction vs

    smooth muscle contraction
- Not all OBO ontologies complete
- No protein / protein complex ontology

Future GO

- Strip out specific instances to leave general concepts in GO
  - eg. metabolism, differentiation, development
- Develop a set of templates for creating composite terms from GO and other OBO ontologies for greater annotation accuracy and flexibility

Future GO

negative regulation of eye photoreceptor cell development

- negative regulation from universal modifier ontology
- eye from anatomy ontology
- photoreceptor cell from cell type ontology
- development from GO process ontology

For more information…

- GO
  - http://www.geneontology.org
- OBO
  - http://obo.sf.net
- Term decomposition / OBOL
  - http://www.fruitfly.org/~cjm/obol-0.02/doc/obol-doc.html