BEL Language 2.0

April 29th, 2014

Natalie Catlett

BEL v2.0 – Coming this summer!

- First revision of OpenBEL
- Planned timeline:
  - Specifications ready in July 2014
  - Compatible with next generation BEL Framework - TBD
- Key areas targeted for update:
  - DNA/RNA/Protein variant representation
  - Translocations and cellular location
  - Activities
- Additional updates:
  - New relationship – *causes*
  - Citation format
- To provide input - please contact me to get involved with the language committee!

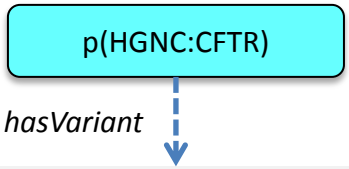
BEL v1.0 Variant Representation

- Key challenges:
  - BEL v1.0 only allows specification of mutations/variants at protein level
    - No means to represent SNPs that do not affect protein coding
  - Requires that both the reference and variant alleles are known and specified
    - No means to *specifically* represent the reference allele
    - No means to represent non-specified mutations
- BEL v1.0 variant representation (protein only)
  - `p(HGNC:APOE, sub(C, 130, R))`
    - APOE protein with cysteine 130 substituted with arginine
  - `p(HGNC:APOE)`
    - "wild-type" APOE protein? Or all APOE protein?
  - `p(HGNC:CFTR, trunc(542))`
    - CFTR protein with truncating nonsense mutation at amino acid 542

BEL Variant Representation Needs

- General goals:
  - Represent variants at DNA, RNA, and protein level
  - Types of variants – substitutions, insertions, deletions, fusions, unspecified, etc.
  - Non-coding DNA variants
  - Equivalencing of representations
  - Representation should support mapping of measurement data to network
  - Compatible with related representation needs– e.g., representation of DNA modifications
- Proposal – adopt HGVS mutation nomenclature for most BEL variation expressions

BEL Variant Representations - Proteins



The root node – CFTR protein abundance
(includes reference allele, and any variants or modifications)

p(HGNC:CFTR, var(=))

The reference allele

p(HGNC:CFTR, var(?))

An unspecified variant

p(REF:NP_000483.3, var(p.Gly576Ala))

CFTR substitution variant *NP_000483.3:p.Gly576Ala*
NOTE – because a specific position is referenced,
an ID for a non-ambiguous sequence is preferred

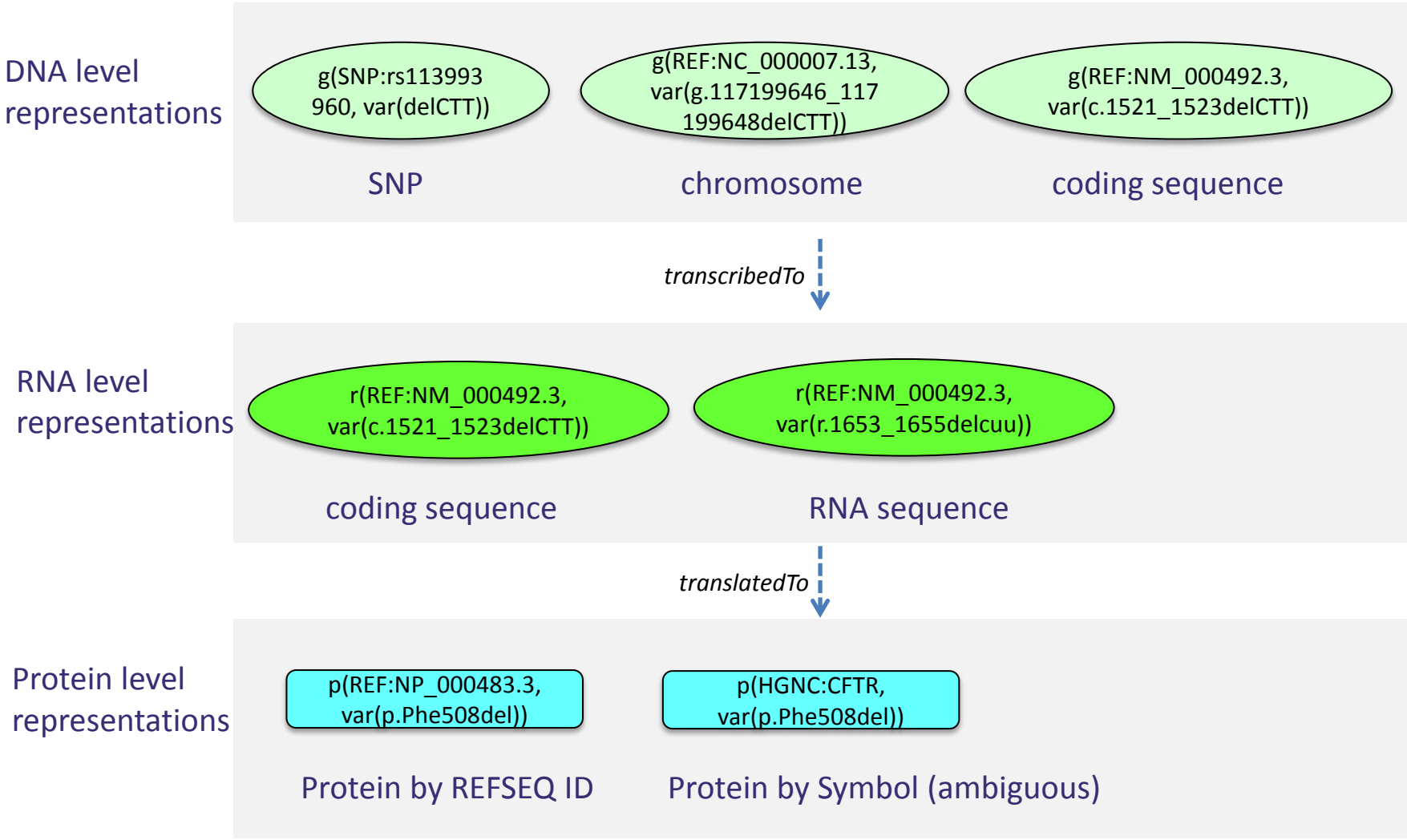
p(REF:NP_000483.3, var(p.Phe508del))

CFTR deletion variant *NP_000483.3:p.Phe508del* (ΔF508)

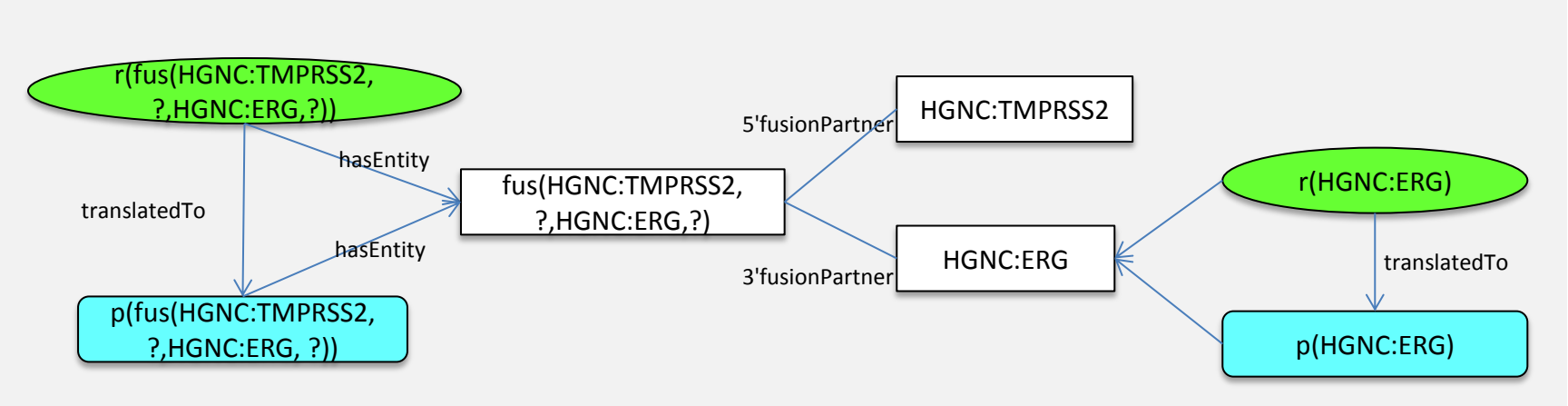
p(REF:NP_000483.3, var(p.Glu726Argfs))

CFTR frameshift variant *NP_000483.3:p.Glu726Argfs*

BEL Variant Representation - DNA and RNA



BEL Variant Representation - Fusions



- Fusions should be represented with a different BEL function than other variants
  - A fusion should link to both the 5'- and 3' fusion partners, but be treated as a "new" gene sequence rather than a variant form of the fusion partners
  - HGVS nomenclature not well documented for fusion representation
- fus(ns:value, range, ns:value, range)
  - Use fusion construct in place of a namespace value for g(), r(), p() BEL functions

Translocation and Cellular Location

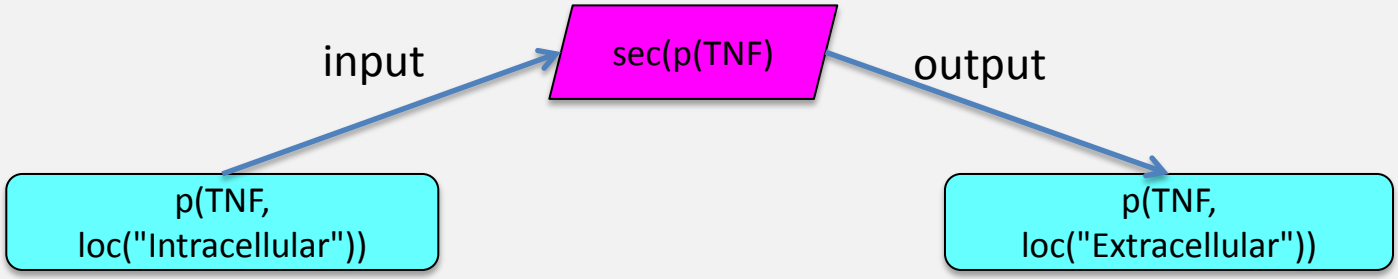
- BEL v1.0 translocation representation:
  - `tloc(p(HGNC:AKT1), MESHCL:Cytoplasm, MESHCL:"Cell Membrane Structures")`
  - Translocation of human AKT1 protein from the cytoplasm to cell membranes
- Challenges:
  - No means to indicate a distinct cellular pool of an abundance, only movement between locations
    - Except via statement annotations
  - Cellular locations used for translocation 'to' and 'from' are not contained within a BEL function
- Proposed solution:
  - Introduce new BEL function `loc()`
  - Represent translocations as transformation of an abundance at one `loc()` to another `loc()`

BEL v2.0 Proposed Translocation and Cellular Location Representation

- Translocation of AKT1 protein from cytoplasm to membranes
  - `tloc(p(HGNC:AKT1), loc(MESHCL:Cytoplasm), loc(MESHCL:"Cell Membrane Structures"))`
- Designation of cytoplasmic pool of AKT1 protein
  - `p(HGNC:AKT1, loc(MESHCL:Cytoplasm))`
  - *loc()* modification should apply to most/all abundance types except *composite() - a(), g(), r(), m(), p(), complex()*

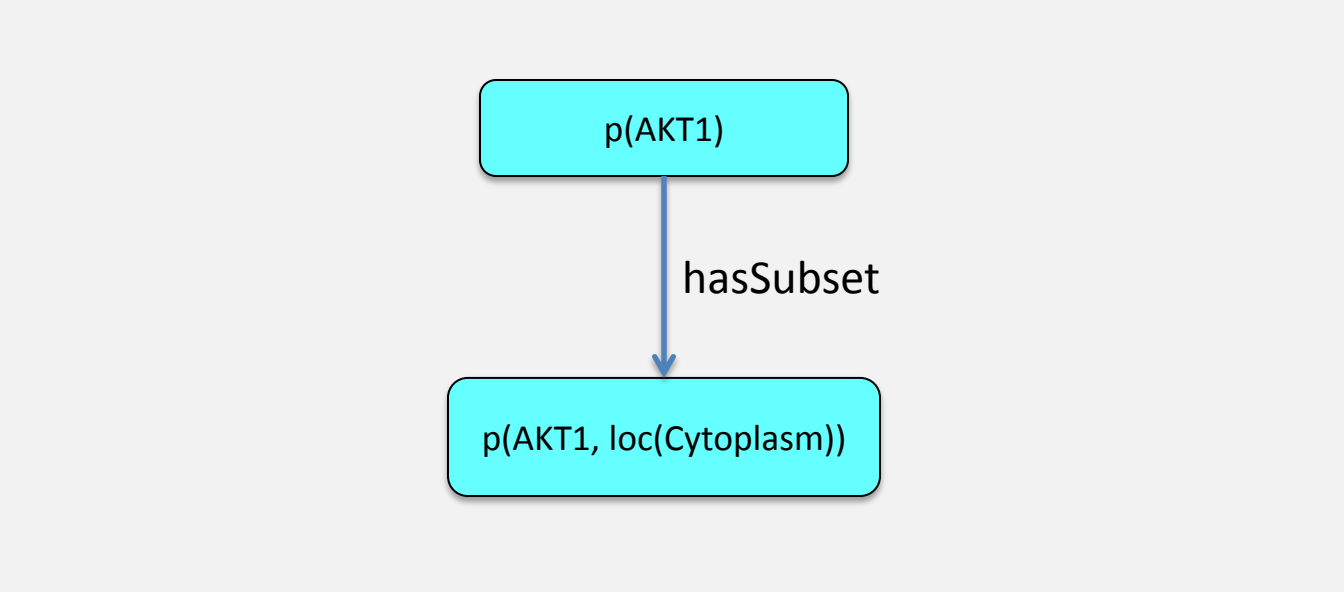
Proposed Compiler Expansions for `translocation()`

- Inferred from translocation term:
  - `sec(p(HGNC:TNF)) output p(HGNC:TNF, loc(MESHCL:"Extracellular Space"))`
  - `p(HGNC:TNF, loc(MESHCL:"Intracellular Space")) input sec(p(HGNC:TNF))`
  - Translocations similar in form to reactions



Proposed Compiler Expansion for `location()`

- Inferred from location term:
  - `p(HGNC:AKT1) hasSubset p(HGNC:AKT1, loc(MESHCL:Cytoplasm))`



Activity Functions

- BEL v1.0 includes 10 activity functions
  - E.g., *kinaseActivity*, *peptidaseActivity*, *transcriptionalActivity*
  - Applied to protein abundances to differentiate the molecular activity from the abundance of a protein
- Challenges
  - Addition of new activities requires a language update
  - Inconsistent usage of specific activity functions
    - Selection of most appropriate activity often requires information outside of the reference text
- Proposed solution
  - Consolidate activity functions to single BEL function ***act()***
  - Capture specific molecular functions via namespace values
    - as a modification of ***act()***
    - Enables use of namespaces for molecular functions

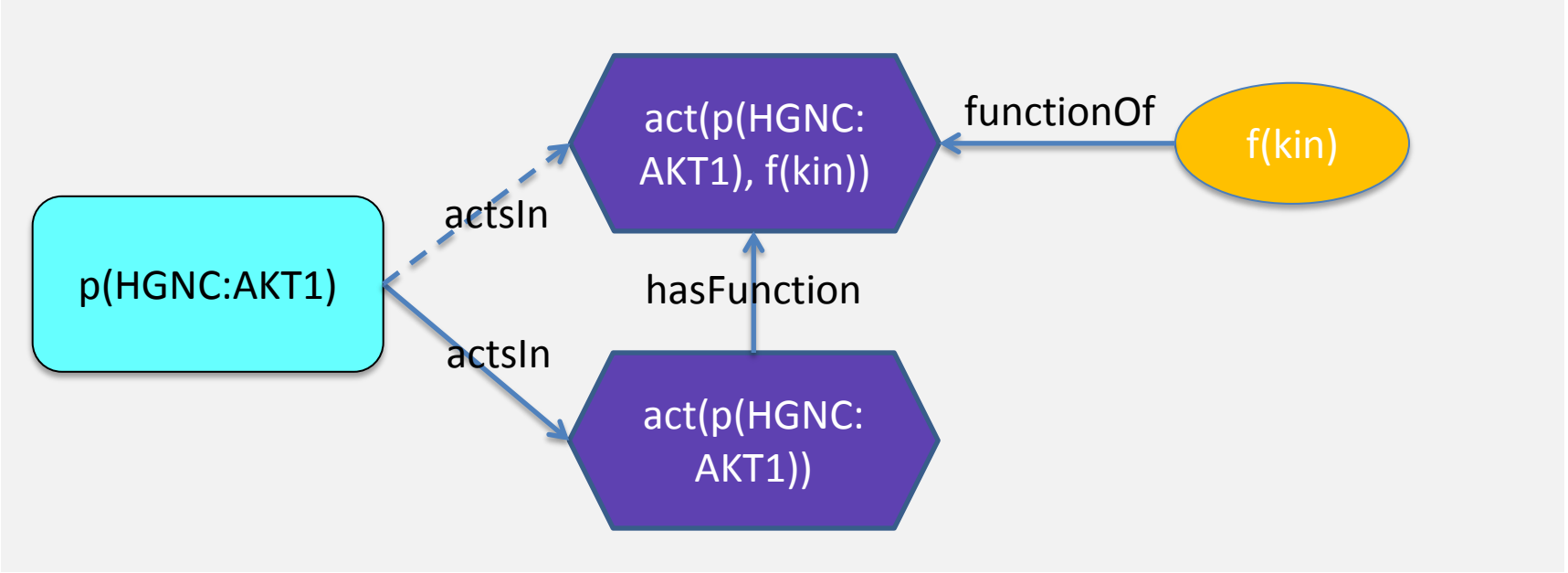
Simplification of BEL Activity Functions to Single Function – act()

Current representation:

Proposed representation:

- Both ***kin()*** and ***tscript()*** functions genericized to ***act()***
- Specific molecular functions can be optionally introduced through ***mf()***

Simplification of BEL Activity Functions to Single Function – *act()*

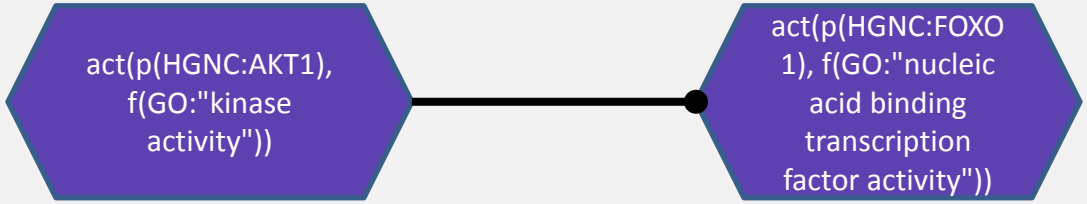


- Activities are connected to activities modified by a function
- Functions are connected to activities

Using GO as a Molecular Function Namespace

Current representation:

Proposed representation:



- Users can introduce any type of activity function
- Both ***kin()*** and ***tscript()*** functions genericized to ***act()***
- Specific molecular functions are introduced through ***f()***

New Relationship - *causes*

- Challenge:
  - Database conversion and text mining project instances where A effects B, but no increase/decrease relationship is specified
  - Lose information about cause and effect by using *association*
- Proposed solution:
  - Add *causes* relationship
  - Collapse with *increase* or *decrease* in KAM
    - Combine edges to use the relationship providing the most information

"Protein A regulates secretion of Protein B"

Citation Annotation Format

- BEL v1.0
  - SET Citation = {"PubMed","Cell","16962653","2006-10-07","Jacinto E|Facchinetti V|Liu D|Soto N|Wei S|Jung SY|Huang Q|Qin J|Su B",""}
  - First three fields (type, name, PMID) are required
  - Supported types are: "Book", "PubMed", "Journal", "Online Reference", "other"
- Challenge:
  - Need easier format with minimal information required to unambiguously identify source

Other High Priority Revisions?

- Expand protein modifications
- DNA modifications