BNS: A DNS-Inspired Biomolecule Naming Service

Robert Kincaid, *Life Science Technologies Laboratory, Agilent Technologies, Palo Alto, California, USA*
robert_kincaid@agilent.com

OVERVIEW

Problem: Biomolecule identifiers lack consistency making informatics tasks more difficult

Maintaining consistency of molecular identifiers, names and annotations over tens of thousands of measurements is an important but daunting task since no standard universal molecular identification scheme exists. For example, associating diverse experimental data derived from both nucleic acid and protein measurements often requires resolving different but related accession numbering schemes. Combining experimental results from different laboratories, or different assay vendors can also require similar effort to correctly associate data labeled according to different identifier schemes. Databases containing the information necessary to address these issues exist, but are generally not amenable to fast programmatic access over tens of thousands of genes at a time, and generally rely on infrastructures unique and often proprietary to the database implementation. Often special single-use ad hoc scripts or programs are written to combine data from different sources. The lack of a ubiquitous general-purpose solution to resolve gene names and identifiers is surprising since this issue is a fundamental component of the larger problems of data integration and federation.

Solution: A biomolecule naming service

Loosely inspired by DNS servers used to resolve host names and IP addresses, we have developed a prototype naming service for biomolecules we call the Biomolecule Naming Service or BNS. This prototype uses the Lightweight Directory Access Protocol (LDAP) to provide high performance, scalable access to data derived from a name- mapping database. We based the prototype on LocusLink† since it is a freely available public source of curated data that contains the necessary identifier and name mappings. However, the architecture can support other data as required. BNS is *specifically* designed and optimized to address the problem of efficiently resolving various name and identifier schemes. Using this system and its associated programming interfaces, it is possible to easily and rapidly resolve gene, transcript and protein identifiers as well as names and aliases into various equivalents.

† RefSeq and LocusLink: NCBI gene-centered resources. Pruitt KD, Maglott DR Nucleic Acids Res 2001 Jan 1;29(1):137-140

BNS PROVIDES STANDARDS-BASED INFRASTRUCURE AND INTERFACES THAT EASILY RESOLVE EXISTING MOLECULAR NAMES AND IDENTIFIERS

BNS is specifically designed to help solve the problem of inconsistent terminology by providing translation between different name and identifier schemes. For example, BNS interfaces have been developed to translate:

- Alias names into official or preferred name
  `PSCP` → `BRCA1`

- RefSeq mRNA accessions to protein counterparts and vice versa
  `NM_007294` → `NP_009225`
  `NP_009226` → `NM_007295`

- RefSeq id's or GenBank identifiers to the underlying locus
  `L78833` → `672`

- RefSeq and GenBank accessions to the underlying UniGene cluster
  `NM_007294` → `Hs.194143`

Facilities for general queries are also provided that allow ad hoc searches against gene ontology, specific organisms and other available entries and attributes. Interfaces have been implemented in Java, C++, Perl and MATLAB in order to support any potential informatics solution.

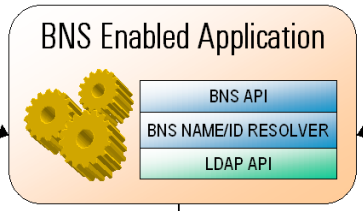
BNS FACILITATES COMBINING DIFFERENT DATA TYPES

Gene Expression Array

UniGene ID	Log Ratio	p-Value
Hs.25647	-1.00655	5.76E-22
Hs.72988	-0.052763	2.34E-19
Hs.50651	0.658436	3.13E-17

Protein Array

Accession	Log Ratio	SDev
NP_005243	-0.947	0.053
NP_005410	0.392	0.012
NP_002218	0.772	0.024



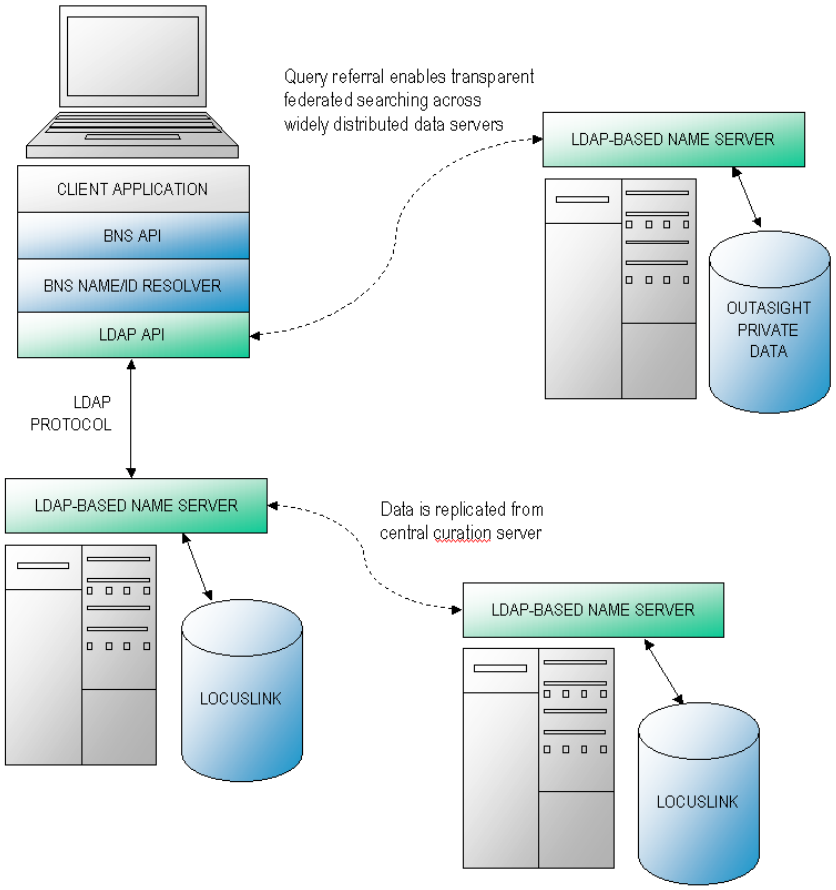
Locus ID	Symbol	Description	UniGene ID	Log Ratio	p-Value	Accession	Log Ratio	SDev
2353	FOS	v-fos FBJ murine osteosar	Hs.25647	-1.00655	5.76E-22	NP_005243	-0.947	0.053
6773	STAT2	signal transducer and acta	Hs.72988	-0.052763	2.34E-19	NP_005410	0.0121	0.012
3716	JAK1	Janus kinase 1 (a protein	Hs.50651	0.658436	3.13E-17	NP_002218	0.772	0.024

One of the initial motivations behind BNS was how to easily combine nucleic acid measurements with protein counterparts. In the figure above we show a hypothetical example combining expression data from a gene expression array (RNA targets) with an antibody array, which produces corresponding protein measurements. The prototype contains a nominal amount of annotation information. While joining the data note that we can also add additional information useful to the interpretation of the measurements.

BNS ALSO ENABLES OTHER USEFUL HIGH-THROUGHPUT TASKS

- **GENERATE ARRAY DESIGNS** – ad hoc BNS queries over descriptive names and/or Gene Ontology entries can be used to create theme-oriented microarray designs. Large annotated gene lists can be generated in minutes.
- **RE-ANNOTATE ARRAY DESIGNS** – we have constructed simple tools to read an existing microarray design (~8K-15K features) and re-annotate the design with the latest Locus Link annotations. This process could be easily automated.
- **CONVERT/NORMALIZE IDENTIFIER SCHEMES** – A data set with a single or mixed set of identifiers can be "normalized" to use the desired common identifier scheme.
- **VERIFY GENE NAMES DURING TEXT MINING** – BNS has been used successfully as a component in text-mining scientific literature. Gene names can be verified and aliases converted to the official gene name.
- **BNS ON-DEMAND** – With just a few lines of code, any application can request detailed information about a given identifier. Information can be presented as a pop-up dialog to provide supplemental information.

QUERY REFERRALS AND DATA REPLICATION PERMIT AN EASY-TO-MAINTAIN DISTRIBUTED NETWORK OF DATA



LDAP supports both ***data replication*** as well as ***query referrals***. Replication permits data to be automatically propagated throughout a network of BNS servers. In this way BNS-stored genomic and proteomic data could in theory become as ubiquitous within bioinformatics systems as Internet addresses are to network software. If BNS infrastructures are widely available, with fairly modest computing infrastructure, a laboratory could have a local high-throughput repository of data that is constantly and automatically updated. One would no longer have to ftp from public servers, extract and parse flat files and reload into local proprietary databases, or subscribe to some commercial service to perform these operations. Further, with query referral, data can be highly optimized and distributed, based on need, performance and security considerations. For example, only data that is likely to be relevant to a particular lab needs to be stored locally. Queries for infrequently used, atypical data can be automatically and transparently referred to shared remote BNS servers.

BNS PROVIDES A SIMPLE PROGRAMMING MODEL WHICH IS EASY TO USE

Here is a short snippet of Java that illustrates how easy it is to use the BNS API:

```
try {
  conn.connect("ldap://bnshost.somedomain.edu");          // STEP 1: Connect to the ldap server
  System.out.println(conn.lookupSymbol("ABL1").toText()); // STEP 2: Do some BNS calls
  conn.disconnect();                                      // STEP 3: Disconnect - All there is to it!
}
catch (BNSException e) {
  e.printStackTrace();
}
```

The resulting output is:

```
LOCUS         25
SYMBOL        ABL1
ALIAS         ABL, JTK7
DESCRIPTION   v-abl Abelson murine leukemia viral oncogene homolog 1
UniGene ID    Hs.146355
GenBank       K00009, M13099, U07563, M14752, M14753, M30833, X16416
TRANSCRIPTS
     NM_005157, NP_005148, v-abl Abelson murine leukemia viral oncogene homolog 1 isoform a
     NM_007313, NP_009297, v-abl Abelson murine leukemia viral oncogene homolog 1 isoform b
GENE ONTOLOGY
     cellular component : 0005634 : nucleus
     biological process : 0007048 : oncogenesis
     molecular function : 0003677 : DNA binding
     biological process : 0006298 : mismatch repair
     biological process : 0000074 : cell cycle control
     biological process : 0006464 : protein modification
     molecular function : 0004713 : protein tyrosine kinase
     biological process : 0006355 : transcription regulation
     biological process : 0000115 : mitotic S-specific transcription
     biological process : 0008630 : induction of apoptosis by DNA damage
CHROMOSOME    9q34.1
SUMMARY       The ABL1 protooncogene encodes a cytoplasmic and nuclear protein tyrosine kinase that has
been implicated in processes of cell differentiation, cell division, cell adhesion, and stress response.
Activity of c-Abl protein is negatively regulated by its SH3 domain, and deletion of the SH3 domain
turns ABL1 into an oncogene.  The t(9;22) translocation results in the head-to-tail fusion of the BCR
(MIM:151410) and ABL1 genes present in many cases of chronic myelogeneous leukemia.  The DNA-binding
activity of the ubiquitously expressed ABL1 tyrosine kinase is regulated by CDC2-mediated
phosphorylation, suggesting a cell cycle function for ABL1.  The ABL1 gene is expressed as either a 6-
or 7-kb mRNA transcript, with alternatively spliced first exons spliced to the common exons 2-11.
```

Getter/Setter functions are provided for all attributes in addition to toText() and toHTML() convenience functions. Further examples are:

`resolveTranscriptionPair("NM_000018")` returns NP_000009.
`resolveTranscriptionPair("NP_000009")` returns NM_00018.
`lookupID(Hs.82208), lookupID(NM_000018), lookupID(NP_000009)` all return locus 37 (ACADVL). lookupID inspects the prefix of the identifier to determine if it's RefSeq or UniGene. Otherwise it assumes GenBank.

CONCLUSIONS AND FUTURE DIRECTIONS

The current BNS prototype (based on LDAP and LocusLink) has sufficient performance and functionality to be useful for a number of tasks common to molecular biologists and bioinformaticians. In particular, it is quite practical to convert molecular identifiers from one system to another during the actual analysis of data (vs. creating completely consistent datasets prior to analysis). Similarly, it is possible to retrieve curated relationships (where available) to associate entities with otherwise unrelated identifier schemes. By using LDAP facilities for query referrals, and replication, it is possible to create highly distributed, largely self-maintaining infrastructures.

While the genome coverage of LocusLink is substantial, further work to expand BNS to integrate identifiers from other public and private databases would improve its usefulness. We have already investigated incorporating licensed, proprietary sequence identifiers into BNS, allowing easy conversion between public and private identifier schemes.

We believe distributed BNS-like systems could become a simple, efficient and effective delivery mechanism for much of the available public genomic and proteomic data. Wider coverage of known genetic loci by LocusLink or equivalents, *combined with a ubiquitous BNS deployment as an open standard*, could provide a powerful new infrastructure useful to molecular biology.

ACKNOWLEDGMENTS

Dean Thompson (LSTL) and Paul Wolber (Bio-Research Solutions) provided encouragement to follow this line of investigation. Aditya Vailaya (LSTL) provided a unique use case by using BNS to assist text mining. Allan Nakagawa, and the members of the Agilent in situ microarray design group as well as Karen Shannon and the members of the Agilent catalog array design group helped test the prototype and provided useful feedback.