

# ginas Software Status

The screenshot displays the ginas software interface, version v0.9301, with a user logged in as Tyler Peryea. The interface includes a navigation bar with 'Home', 'Search', 'Register', and 'Download' links. The main workspace is titled 'Nucleic Acid Sequence:' and shows a sequence of 1 subunits. A chemical structure overlay is visible, showing a phosphate group (P) with a double bond to an oxygen (O) and a single bond to a hydroxyl group (OH). The phosphate is connected to a 5' carbon (orange) and a 3' carbon (green). The sequence editor shows a sequence of nucleotides: A-A-A-T-C-T-A-T-C-T-A-T-C-G-G-A-C-T-A-G. The sequence is displayed in a box with a 'C' at the 5'-end and a 'G' at the 3'-end. The interface also includes buttons for 'Previous', 'Cancel', 'Next', 'Edit Sequence', 'Sugars', 'Linkages', 'Modifications', and 'Nucleic Acid Type'.

Dac-Trung Nguyen, Tyler Peryea

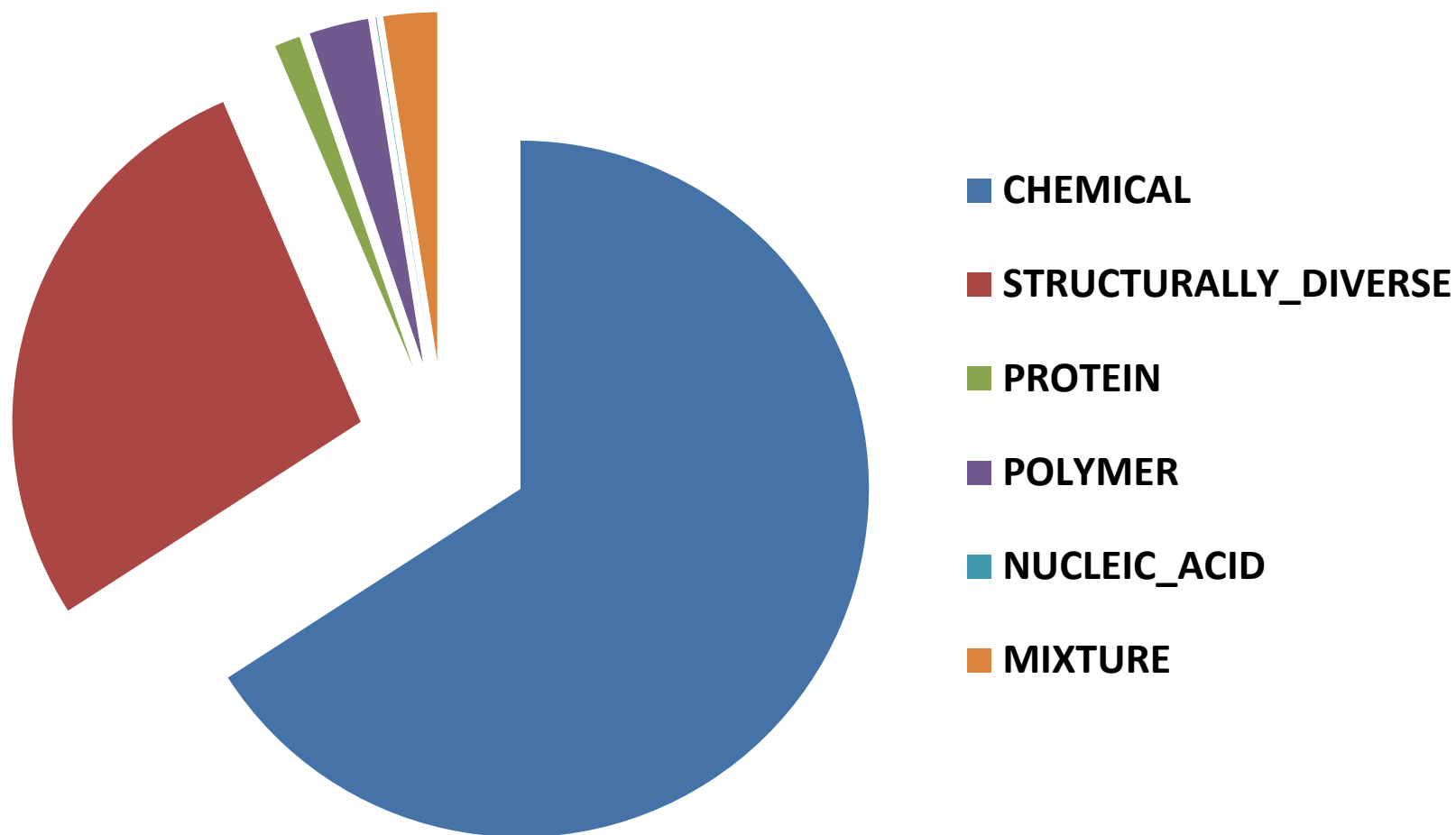
NCATS



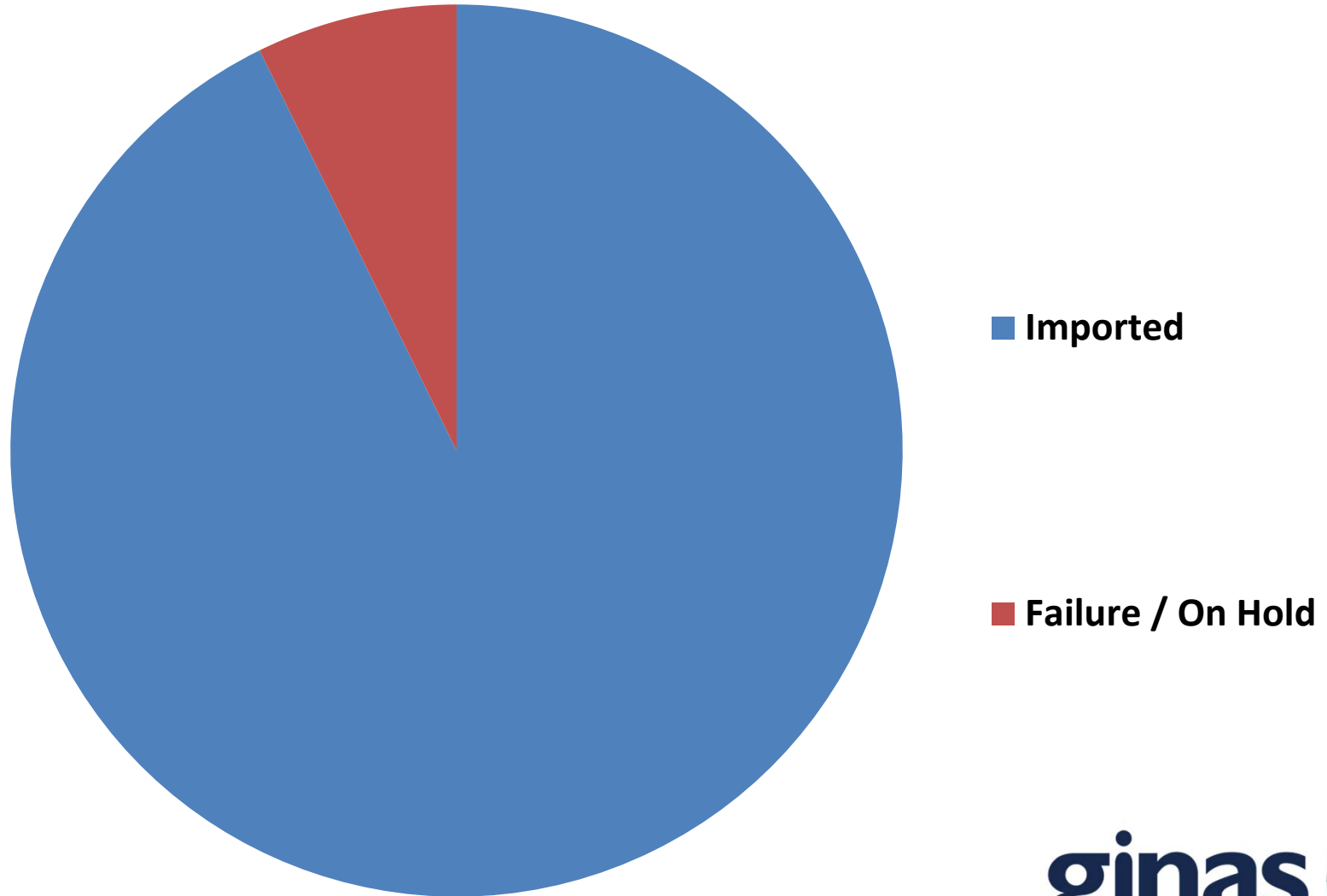
# Software Status

- Registration Coverage
  - Web interface and data import path
- Global-ness
- Distributable-ness
- Usability
- Security and User management
- Data Quality and Data Management
- Data Exchange
- “Open Source” status
- Links

# SRS Distribution of Substance classes



# SRS Transformation Count



# Software Status : Registration

	Conceptual Case Study	Implemented	Live Case Study	Large Scale Import Study
<b>Protein</b>	YES	YES	YES	In process
<b>Chemical</b>	YES	YES	YES	In process
<b>Nucleic Acid</b>	YES	YES	YES	In process
<b>Mixture</b>	YES	YES	YES	In process
<b>Structurally Diverse</b>	YES	YES	YES	In process
<b>Polymer</b>	YES	YES	YES	In process
<b>G1SS</b>	YES	YES	YES	NO
<b>G2SS</b>	NO	NO	NO	NO
<b>G3SS</b>	NO	NO	NO	NO
<b>G4SS</b>	NO	NO	NO	NO

# SRS Import Tests

- Developing adapter from SRS format to ginas format, which is run periodically
- Some things fail due to the adapter, some things fail due to the data

## June 6<sup>th</sup> Status

Status	Count	Meaning
UNEVALUATED	326	Not yet attempted or fundamental I/O problem
LOADED	317	Basic table info assembled from SRS tables, but no XML/structural parsing possible
PARSED	807	Description parsed correctly, but no adapter yet, or in unexpected/invalid state
ADAPTED	3436	Light-version of GINAS adaptation / validation
PROCESSED	1766	Heavy-version of GINAS validation / formatting
SUBMITTED	56150	Successfully entered into a GINAS instance
TOTAL	<b>62802</b>	89%

CLASS	Total	Submitted	Preliminary Percentage
CHEMICAL	40937	37487	92%
STRUCTURALLY_DIVERSE	17189	17051	99%
PROTEIN	744	508	68%
POLYMER	1708	1469	86%
NUCLEIC_ACID	22	0	0%
MIXTURE	1542	1103	72%
Total	62142	57618	93%



# Software Status : **Global-ness**

- Full Unicode support for entry/searching across many different languages
- Translations for *most* INN names now present for 6 languages:
  - Russian
  - Spanish
  - English
  - Chinese
  - Arabic
  - French

# Software Status : **Global-ness**

- To do:
  - Regional translations for software text
  - Regional translations for controlled vocabulary
  - Regional preferences for naming display

# Software Status : **Distributable-ness**

- Embedded instance now compiled:
  - Self-contained H2 Database
  - Self-contained Java Web Servlet
  - Bootable image for CD/USB stick
    - Live-boot GNU/Linux (slax) with preset configuration
    - Portable virtual machine
  - Pre-compiled with preliminary sample data



# Software Status : **Distributable-ness**

- To do:
  - Detailed documentation of specific setup
  - Optimize initialization
  - Optimize memory footprint
  - Procedure for updates
    - For data
    - For software

# Software Status : **Use-ability**

- Many convenience tools for searching and registration:
  - Copy/paste chemical structure browser plugin
  - Image-to-structure browser plugin
  - “Draw-ahead” substructure searching
  - Real-time validation and feedback in many areas
  - REST API excel plugin for quickly resolving names/structures to ginas ID
  - Sequence homology searching (proteins and nucleic acids)

# Software Status : Use-ability

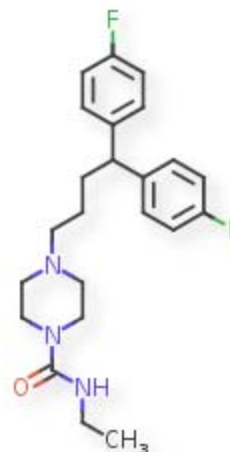
- Excel example

2				
3	DIETHANOLAMINE BISULFATE			
4	MORPHOCYCLINE			
5	QUINUCLIIUM BROMIDE ANHYDROUS			
6	ISOTIQUIMIDE			
7	CYSTEAMINE			
8	AMPEROZIDE			
9	IMAZAMOX-AMMONIUM			
10	SODIUM MYRISTYL SULFATE			
11	ORG-28611 HYDROCHLORIDE			
12	NOR-URSODEOXYCHOLIC ACID			
13	CARBOFURAN			
14				
15				
16				
17				

# Software Status : Use-ability

- Excel example

2		
3	DIETHANOLAMINE BISULFATE	OS(O)(=O)FDA-SRS
4	MORPHOCYCLINE	CN(C)[C@FDA-SRS
5	QUINUCLIU BROMIDE ANHYDROUS	[Br-].C[N+FDA-SRS
6	ISOTIQUIMIDE	CC1=CC=NFDA-SRS
7	CYSTEAMINE	NCCS FDA-SRS
8	AMPEROZIDE	CCNC(=O)FDA-SRS
9	IMAZAMOX-AMMONIUM	[NH4+].CCP FDA-SRS
10	SODIUM MYRISTYL SULFATE	[Na+].CCCCFDA-SRS
11	ORG-28611 HYDROCHLORIDE	
12	NOR-URSODEOXYCHOLIC ACID	C[C@H](C FDA-SRS
13	CARBOFURAN	CNC(=O)CFD
14		
15		
16		
17		
18		
19		
20		
21		
22		



# Software Status : **Use-ability**

- To Do:
  - Improve registration step-through wizards
    - (particularly in polymers, structurally diverse)
  - Improve record display for browse-ability rather than strict data elements
  - Improve searching/browsing and filtering capabilities
  - Implement export procedures to commonly used formats (sdf, excel, etc)



# Software Status :

## **Security and User management**

- Users allowed with various roles
- Roles control what that user is allowed to do
  - Register
  - Approve
  - View
  - Update
- Done now as static users, inherent to embedded system

# Software Status :

## **Security and User management**

- To Do:
  - All public releases have only public data, so deeper security model has been triaged
  - Authentication and private/public key encryption for all information sent
  - Database-level security implementation for embedded and production system

# Software Status :

## Data Quality and Data Management

- REST API in place to query and return full substance object (in JSON format)
- Fairly static format for objects
- Timestamps, owners, and references for every piece of information
- Very simple duplication detection for chemicals

# Software Status :

## Data Quality and Data Management

- To Do:
  - Allow for easy backend SQL querying inherent to model
  - Model for each substance class must be more solidified in a few areas (cardinality issues, etc)
  - Controlled Vocabularies preliminary, and must be re-evaluated, and pointed to external authorities where available (e.g. Kew Gardens)
  - Proper “fuzzy” duplication detection for all substance classes
  - REST API exposure of change log and versioning

# Software Status :

## Data Exchange

- Object structure can be readily exported / imported into different ginas instances
- To Do:
  - Common exchange mechanism (semi-automatic imports into other systems)
  - Merging / conflict resolution reporting on exchange

# Software Status :

## **“Open source” Status**

- Mostly open source software, with a few licensed, distributable commercial packages
- Code available on private NCATS GitHub account
  - Limited number of seats
- NCATS-specific git repository will be available soon for all who request access
- REST API documentation available on github wiki
- To Do:
  - Transition to public completely public git repository
  - Publish API specifications for other developers

# Links:

- <http://ginas.hc.ircan-rican.org/ginas/>
- <http://tripod.nih.gov/ginas/>

The screenshot displays the GINAS web application interface. At the top, the logo "ginas" is shown next to a circular icon with three vertical bars. Below the logo, the text "development v0.9301" is visible. To the right of the logo, there is a language selector set to "U.S.A.", a login section with fields for "username" and "password", and a "Log In" button. A navigation bar contains links for "Home", "Search", "Register", "Download", and "Report a Bug", followed by a search input field labeled "search...".

On the left side, a sidebar contains a "Show All" button with a count of 30, and two sub-sections: "Substances" with a count of 30, and "Specified Substances" with a count of 0. Below these, a "Substance Types (1)" section shows "CHEMICAL" with a count of 30. A "Query" section displays a chemical structure of toluene (a benzene ring with a methyl group, CC1=CC=CC=C1).

The main content area displays a grid of search results, each showing a chemical structure and its similarity score to the query. The results are as follows:

Similarity	Chemical Structure	Name
Similarity: 0.91	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	1-METHYLNAPHTHALENE
Similarity: 0.91	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	2-METHYLNAPHTHALENE
Similarity: 0.89	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	1,2-DIMETHYLNAPHTHAL...
Similarity: 0.89	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	3-METHYLPHENANTHRENE
Similarity: 0.86	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	
Similarity: 0.54	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	
Similarity: 0.82	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	
Similarity: 0.65	<chem>CC1=CC=C2C(=C1)C=CC=C2</chem>	