

Background on the PubChem/CAS Issue

What is the issue?

The American Chemical Society (ACS) has expressed concern that a new NIH database called PubChem, which was created to support the NIH Roadmap initiative, and especially the Molecular Libraries Initiative, is a threat to the financial survival of the Chemical Abstracts Service (CAS). PubChem provides free access to its database; CAS charges a fee for researchers to use its database. ACS has demanded that NIH shut down PubChem or substantially alter it so as not to compete with CAS.

What is PubChem?

In 2004, as part of the NIH's Roadmap Initiative to speed new medical treatments and improved health care to all Americans, NIH launched an on-line database called PubChem as part of a suite of databases supporting the New Pathways to Discovery component of the roadmap effort. New Pathways focuses on very basic biomedical research, and especially focuses on understanding the molecular biology of health and illnesses. Bioinformatics is a critical component of that effort.

Drawing from many public sources, PubChem organizes information about the biological activities of chemical compounds into a comprehensive biomedical database. All of this supports the part of the Roadmap called the Molecular Libraries initiative. This includes nine different components – a compound repository, the NIH Chemical Genomics Center, the Molecular Libraries Screening Center Network, PubChem, a series of Cheminformatics Research Centers, and technology development for chemical diversity synthesis, assay development, instrumentation, and toxicology. PubChem is the informatics backbone for virtually all of these components, and is intended to empower the scientific community to use small molecule chemical compounds in their research.

Without PubChem, the work of NIH funded scientists will be greatly hampered and progress in biomedical research will be slowed.

Why is ACS/CAS concerned?

Information gathered through conversations with ACS officials and media accounts indicates that:

1. ACS/CAS is concerned that a federally supported database that is freely available to all users and is supported by federal tax dollars has an unfair advantage over the CAS service, which charges a fee for access to its database.
2. ACS claims that PubChem will cripple CAS, sapping both it and the ACS's economic foundation, resulting in the loss of jobs in Columbus, Ohio.
3. ACS/CAS appears to want NIH to either shut down the PubChem database or severely limit its content so that it does not overlap with ACS/CAS in any way.

4. ACS/CAS also appears to want to provide some of the information contained in PubChem, but at a cost to researchers who would use this information.

How does NIH respond to ACS's concerns?

1. PubChem is an essential component of a major NIH Roadmap Initiative to advance medical research.
2. PubChem is part of an integrated suite of NIH databases and these cross connections are critical to optimal functionality.
3. NIH believes that PubChem and the Chemical Abstracts Service databases are complementary and not duplicative.
4. NIH believes that CAS is a valuable resource to chemists – and could be a valuable resource to biomedical researchers, who currently do not focus on the information organized in CAS. NIH believes that will be achieved through collaboration with CAS.

Similarities and Differences between PubChem and CAS:

NIH staff conducted an analysis to compare PubChem and CAS. A number of important findings emerged:

1. Staffing:
 - a. CAS employs approximately 1,300 staff to manually curate the Chemical Abstracts Service databases and information sources.
 - b. NIH has assigned 13 staff to work on PubChem.
2. Budget:
 - a. CAS is reported to have an annual budget of \$260 million, accounting for more than half of the \$421 million budget of the American Chemical Society (according to Business Week, 4-25-05).
 - b. NIH has budgeted \$3 million annually for PubChem.
3. Scope of information
 - a. CAS contains detailed information on approximately 25 million unique chemicals.
 - b. PubChem contains rudimentary information on 650,000 unique chemicals (though this number is expected to grow).

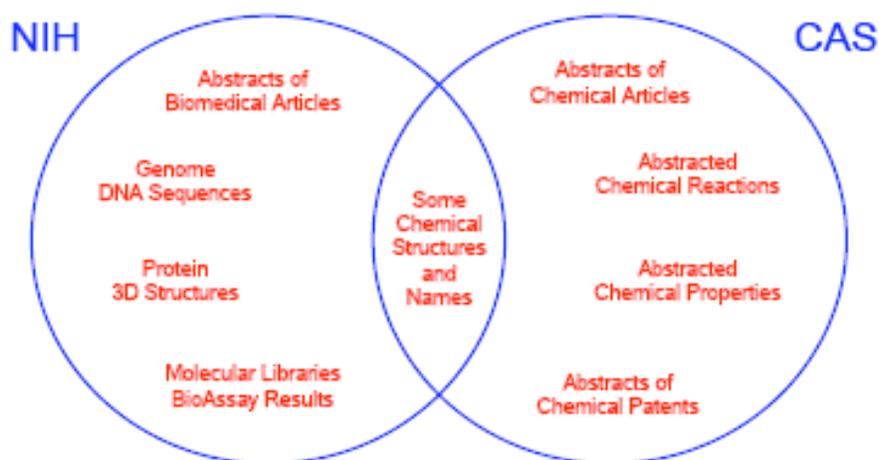
4. Overlap of PubChem/CAS information

On a statistical basis, the actual overlap between the information in the PubChem database and the CAS database is actually quite modest, and for the most part complementary, as shown in the following table and Venn diagram:

Information Content Comparison ...

	NIH	CAS
Chemical Structures	Automated Extraction from Publicly Available Sources	Manual Abstraction by Experts from Chemical Literature and Patents
Measured Chemical Properties	None	Manual Abstraction by Experts from Chemical Literature and Patents
Patent Information	None	Extensive Summaries Written and Curated by Experts
Research Literature (CAS Core Journals Overlap PubMed by ~30%)	Extensive links to PubMed: Biomedical Literature Indexed by Experts	Extensive links to Chemical Literature Indexed by Experts (CAS includes PubMed but Retrieval Differs)
Chemical Reactions (for Chemical Synthesis, etc.)	None	Extensive Collection Curated by Experts
Biological Assay Results	Yes	No
Integration with Protein 3D Structure	Yes	No

Information Overlap ...



... Minimal

It is also instructive to look at examples that compare the information contained in PubChem and the CAS database. One comparison, shown in part (a), uses a chemical that CAS has chosen to highlight within its presentation on PubChem. The other, shown in part (b), is a chemical selected by NIH as representative of the kinds of chemical information PubChem would typically contain.

- a. CAS chemical: RDX, an explosive.
 - i. Found in both databases.
 - ii. NIH information includes bioassay results, toxicology, biomedical vocabulary, pharmacology and links to the biomedical literature.
 - iii. AS information includes chemical properties, chemical literature, chemical patents and chemical reactions.
 - iv. Te information appears to be complementary, as is shown in the following table.

Summary of NIH and CAS Information
for Example "RDX" ...

	NIH	CAS
Total Articles	140	5,242
Subjects	Biomedical	Chemical
BioAssays	Yes	No
Patents	No	Yes
Reactions	No	Yes
Measured Properties	No	Yes

- b. NIH chemical: fluorouracil, an anticancer drug
 - i. Found in both databases.
 - ii. NIH information includes bioassay results, three-dimensional protein structure, toxicology, biomedical vocabulary, biomedical summary, pharmacology (anticancer activity), toxicology, and links to the biomedical literature.
 - iii. CAS information includes chemical properties, chemical literature, chemical patents and chemical reactions.
 - iv. The information appears to be complementary, as is shown in the following table.

Summary of NIH and CAS Information
for Example "Fluorouracil" ...

	NIH	CAS
Total Articles	21,579	14,435
Subjects	Biomedical	Chemical and Biomedical
BioAssays	Yes	No
Protein 3D	Yes	No
Reactions	No	Yes
Patents	No	Yes
Measured Properties	No	Yes

5. Throughout this discussion, the comparison has been made between PubChem and CAS. The fact is, however, that the American Chemical Society does not make the Chemical Abstracts Service compound registry available as a discrete product. The compound registry is included in the suite of services provided by SciFinder, CAS's most popular internet retrieval product, and in other CAS on-line products. This makes it actually difficult to directly compare PubChem to the ACS/CAS services because what ACS/CAS sells is actually much larger than the compound registry alone.

A Win-Win Solution: Linking PubChem to the Chemical Abstracts Service

1. To provide PubChem users with links to the more detailed chemical information contained in CAS, NIH would like to provide links to CAS in the PubChem database. For this type of linkage to work effectively, however, NIH and CAS would need to work together to link the databases. This is possible only if CAS wishes to cooperate to achieve this goal.
2. NIH believes that linking PubChem to CAS would actually benefit CAS by bringing new customers from the biological community to use its services, customers who currently do not use the CAS database.

NIH's Position

1. NIH wants to continue to operate PubChem as a free, publicly available and integral part of the NIH Roadmap Initiative. This is consistent with the principles of publicly funded science.
2. NIH wants to speed up research to find solutions to America's most important health problems through the development of new therapeutics. PubChem, as part of a suite of databases in the Pathways to Discovery component of the NIH Roadmap, is a critical part of that effort.
3. NIH wants to find a way to collaborate with ACS/CAS in such a way that maintains the fundamental principle of making the results of publicly funded research freely available.
4. NIH wants to find a way for both ACS and NIH to run their respective databases for the benefit of the entire scientific community and the public that supports it.