Hyperactive Molecules and the World-Wide-Web Information System†‡

Omer Casher,[a] **Gudge K. Chandramohan,**[b] **Martin J. Hargreaves,**[b] **Christopher Leach,**[a]
Peter Murray-Rust,[b] **Henry S. Rzepa,**[*,a] **Roger Sayle**[b] **and Benjamin J. Whitaker**[*,c]
[a] *Department of Chemistry, Imperial College of Science Technology and Medicine, London, UK SW7 2AY*
[b] *Chemistry Research Division, Glaxo Research and Development, Greenford, Middlesex, UK UB6 0HE*
[c] *School of Chemistry, University of Leeds, Leeds, UK LS2 9JT*

We propose an Internet based standard based on primary chemical MIME types to be used in conjuction with the World-Wide-Web information delivery system. We show how this allows the user to have access to and control of molecular structural and spectroscopic information transparently integrated into scientific publications. We illustrate this with examples of what we have termed 'hyperactive molecules' from the areas of quantitative structural modelling, organic synthesis, crystallography, molecular dynamics, NMR, mass and laser emission spectroscopy. The on-screen annotation of 3D molecular diagrams is achieved using a Chemical Structure Markup Language (CSML) and is illustrated with hyperlinks embedded in *e.g.* two dimensional diagrams in NMR spectra. The implications of these mechanisms for scientific publication in general, the impact on the quality and reproducibility of published experimental data and the enhancement of serendipitous discoveries are discussed.

The development of molecular coordinate and connectivity information depositories such as the Cambridge structural data base[1] or the protein data bank[2] has been greatly facilitated during the last two decades by computer and in particular network technology. Even so, 12 months or more can often elapse between the acquisition of such data and the dissemination of this information in the form of a published research paper. There are often further delays whilst the information is abstracted and entered into the database, and the latter made available to users. Its local use may require acquisition of a proprietary or a commercial program, or require knowledge of a particular search syntax.

More fundamentally, by far the greater bulk of what is referred to as 'supplemental data' is never made available to the scientific community at large. In the area of molecular modelling, the lack of a full specification of the parameters involved in any specific calculation makes many calculations effectively irreproducible. For example, molecular force fields are rarely quoted in full and the final molecular coordinates derived from the minimization of large biomolecules are equally unavailable. Much original spectroscopic data is also 'lost' after the original analysis is performed by authors. The reasons for this are obvious. Traditional methods of information dissemination, *i.e.* printed journals and books, simply cannot afford the space to publish the source data. Furthermore, the possibilities for transcription errors are likely to make these data useless even if the space could be found to print them.

The recent development of the World-Wide-Web, a com-

Table 1 Anonymous ftp sites for obtaining WWW browsers and tools

ftp.ncsa.uiuc.edu/Web	NCSA Mosaic and tools
ftp.einet.com	MacWeb and WinWeb
ftp.mcom.com/pub	Mosaic Netscape
ftp.dcs.ed.ac.uk/pub/rasmol/	RasMol Molecule Viewer
ftp.ch.ic.ac.uk/pub/eyechem	Explorer EyeChem
ftp.ch.ic.ac.uk/pub/csml	The CSML markup script

munications system based on the global Internet network, offers an interesting solution to at least some of these problems.[3] This mechanism also addresses the mismatch between conventional two dimensional printed pages and the three dimensional properties of most molecules, a particulary acute problem for molecules of biological importance. In this paper, we describe the implementation of what we have called the 'hyperactive molecule concept' into the World-Wide-Web infra-structure, and illustrate its use with biologically oriented examples drawn from different areas of structural, spectroscopic and computational chemistry.

Computational Methodology

The World-Wide-Web operates a communication protocol known as HTTP (hypertext-transport-protocol) and a text and graphical markup language known as HTML (hypertext-markup-language). Within this structure, markup commands known as HTML tags are used to define hyperlinks within a document identifying the Internet location of other documents or molecular information. It is the task of the HTTP protocol to retrieve these remote document fragments and of HTML to display and assemble the contents in an appropriate and consistent manner on the computer screen. Programs that implement both the HTTP and the HTML protocols are known as WWW Browsers. Table 1 contains 'anonymous' ftp sites that can be used to acquire no-cost versions of various browsers and the associated visualization tools. Demonstrations of the various concepts discussed in this paper can be achieved using these programs.[4]

Using this mechanism a molecule referred to in a document can be associated with a set of molecular coordinates. The coordinate data themselves need not necessarily be stored on

---

† *Editor's footnote.* This article is being published in the hope that it will further stimulate discussion and encourage the use of wide-area electronic networks in organic chemical research. Correspondence concerning the subject matter should be addressed to Dr. H. S. Rzepa (e-mail rzepa@ic.ac.uk) or Dr. B. J. Whitaker (e-mail benw@chem.leeds.ac.uk).

‡ *Author's footnote.* This paper is the full version of an earlier communication (ref. 3). Whilst this paper is unconventional in not presenting new experimental or theoretical chemical results, we believe it lays the basis for a revolution in the area of information delivery and chemical communication. We also fully expect that entirely new research tools will emerge from the technologies described here, as well as resulting in an improvement in the delivery and quality of numerical and visual information.

the same machine as the document itself since the remote information is associated with a unique name on the Internet that enables it to be almost instantly located. Currently, the scheme in operation uses what are called URLs or Uniform Resource Locators, which specify a precise physical location for the information. It has been proposed that[5] this naming formalism will be augmented by the more robust URN or Uniform Resource Name, which would allow relocation of the information in a manner transparent to the user. Use of these labels allows a close association between the discussion of molecular properties in a scientific paper and any further information located on a remote Internet based system. Such a hyperlink when activated can result in transfer of e.g. molecular coordinates to the reader of the paper, followed if required by the initiation of some appropriate visualization program. Such a mechanism applies equally well to spectroscopic and other forms of numerical or visual data.

Chemical MIME Types.—Before such a mechanism can be implemented, some convention must be adopted for identifying the precise format in which the remote data is defined and how the data will be processed once it has been transferred locally. We have chosen to adopt the MIME formalism, or Multipurpose Internet Mail Extension.[6] This specifies a primary MIME type, which as currently formulated can be any of seven types; text, application, image, audio, video, message or multipart. These are qualified with secondary types. For example, the primary text type can be qualified by secondary types of plain, richtext, postscript, html amongst others. Thus a MIME specification of 'text/html' would unambiguously indicate a document which can be read using a HTML compliant program. The existing MIME types however, do not readily correspond to the variety of chemical information available and in order to address this we have proposed in an Internet draft[7] that a new primary MIME type to be called *chemical* be defined along with a variety of secondary types. In order that this proposal satisfy the requirements for Internet ratification, all the secondary types must follow definitions that are readily available in the literature. It is permissible to use non-standard types, but these must be clearly indicated with the prefix of an *x* before the secondary type name.

In this paper, we illustrate a number of our proposed MIME types with a number of working examples and using freely available WWW browsers (Table 1). To do this, our proposed MIME types have been configured into the two public World-Wide-Web HTTP servers and are accessible *via* the URL definitions http://www.ch.ic.ac.uk/chemical_mime.html and http://chem.leeds.ac.uk/Project/MIME.html. The molecular information is physically stored on either of these systems and the MIME type is identified simply from the suffix given to the file names. At the reader's end, a World-Wide-Web client program (a browser) is configured with an identical set of MIME types, each of which is then associated by the user with a computer program capable of displaying the content in suitable form. The last action is entirely at the discretion of the user, who may wish to use an expensive commercial program to visualize molecular information, or who may simply wish to view it using a simple (and free) text editor. The specific details of how to associate a program with specific MIME types are dependent on the type of computer used. On most Unix workstations for example, a file with the name *.mailcap* is created in the users directory with entries of the type chemical/x-pdb; rasmol %s, which indicates that a coordinate file identified with the secondary MIME type *x-pdb* will be passed to the RasMol[8] program, which is presumed resident on the user's system.

In Table 2, we list some of the secondary MIME types that have been implemented on our HTTP servers, together with the

Table 2 Chemical MIME types used for hyperactive molecule display

Proposed MIME type	Associated applications
chemical/x-pbd	RasMol, XMol, EyeChem, Ball-and-Stick
chemical/x-mopac	XMol, EyeChem, Ball-and-Stick, Chem3D
chemical/x-gaussian	XMol, EyeChem
chemical/x-eye2eye	EyeChem
chemical/x-csml	CSML Markup file for RasMol

application programs with which we have tested the concept. The secondary MIME types x-pdb, x-mopac and x-gaussian are fairly obviously associated with protein database, MOPAC and Gaussian format files (the *x* prefix indicates that formal Internet ratification is not yet achieved and will be removed once this is accomplished). The x-eye2eye secondary MIME type is associated with the EyeChem[9] module suite written for Irix Explorer and allows two users, at spatially separated sites, to share and manipulate simultaneously a 3D representation of molecular coordinate data. This concept is discussed in more detail below. In implementing application programs we concentrated on the use of high performance workstations (SGI IndySC and IPC), but have also evaluated Macintosh and MS-Windows systems. Although the graphical performance is inferior on the latter systems, the concept is entirely viable on all the types of computer with which we have worked.

Results and Discussion

In this paper we illustrate the concept of chemical MIME types with a number of examples. We focus in particular on the context of interpretation of spectral and conformational data. We wish to emphasize, however, that the concept of a chemical MIME type is generic to chemically related information and, indeed, we have demonstrated elsewhere[3] a number of other applications. The conventional paper based medium of this Journal has certain difficulties when it comes to demonstrating these concepts in action. The URL references given above should therefore be consulted for 'working' versions of the examples discussed below.

Hyperactive Proteins.—The first example relates to the proposed protein database format (pdb), which is an accepted standard data format for the repesentation of protein structures amongst chemists and molecular biologists. Coordinate information of this kind is virtually incomprehensible without some form of graphical visualization, but fortunately there are many excellent graphical programs that support this format. In our implementation, a two-tier system of hyperlinks has been adopted. The first hyperlink invokes a small bit-mapped graphical image or thumbnail sketch of the molecule, which serves to identify the nature of the molecule and assists as a browsing mechanism to attract readers' interest. Such images are normally relatively small computer files and do not take much time to transfer across the Internet network. This icon is itself then linked to the pdb coordinate file, which is transferred to the local system using the *chemical/x-pdb* MIME definition and passed as input data to a visualizer (Fig. 1).

In principle any suitable application program could be associated with the MIME type *chemical/x-pdb*, but we have specifically investigated the use of RasMol, XMol[10] and Eye-Chem, which are all easily obtainable by anonymous ftp over the Internet. These programs in turn allow the user to change the viewing position of the coordinates, to add other features such as a ribbon representation and to further manipulate the data. In principle, the data could be passed[11] to a computationally based program such as e.g. MOPAC,[12] where appropriate calculations derived from the coordinates could be made.
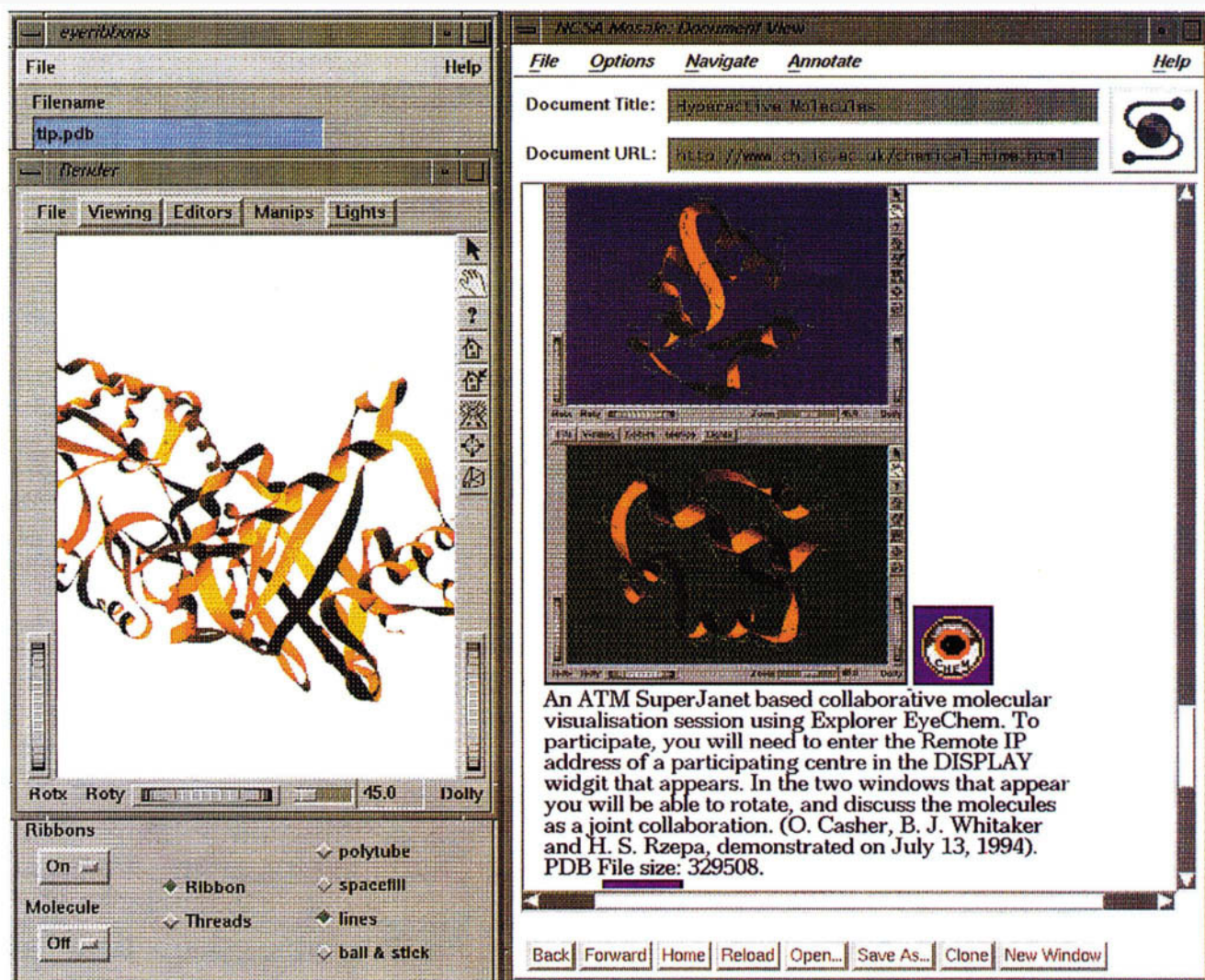
Fig. 1 Triose phosphate isomerase, illustrating the use of the EyeChem viewer in conjuction with hyperlinks inserted into text based references

In this way, the concept of a 'hyperactive molecule' greatly enhances the value of the original document. It also serves as a powerful quality control mechanism, since the scientific integrity of the data can be assessed both at the refereeing stage and subsequently, without the difficulties that are associated with the present mechanism of waiting for the coordinates to be deposited in a single world-wide database.

Quantum Chemically Derived Data.—The concept can be extended to input and output files derived from MOPAC and Gaussian 92 calculations.[13] The visualizers we used were XMol and EyeChem on the X-Windows systems and CAChe, Ball-and-Stick and Chem3D on Macintosh computers. The hyperactive molecules allow the results of the original calculations to be visualized and enable these calculations to be readily repeated if necessary. If a new molecular property is desired, it could be calculated on-the-fly in the context of reading (or indeed refereeing) the original paper. A specific application of this is the annotation of a conformational energy map[14] for the rotation of the two hydroxy groups in the di-axial conformation of *cis*-cyclohexane-1,3-diol. Particular regions of the isometric projection of this map are hyperlinked to molecular coordinates stored on the Web server. By clicking with the cursor on these hot-spots, the relevant coordinates can be acquired and the particular orientation of the hydroxy groups as a function of the calculated energy verified (Fig. 2).

Calculation output files often contain a rich body of information that is rarely published. Often, the view is taken that it is actually easier to repeat the original calculation from the start than to try to acquire such supplemental information. We have demonstrated that such output files, if identified with an appropriate MIME type, can be readily retrieved and processed. For example, a MOPAC force calculation can be retrieved and a particular normal vibrational mode animated to reveal the form of the vibration. A Gaussian log file corresponding to 15 cycles of a geometry optimization can be displayed in animated form. For both these visualizations, we used the EyeChem package, which allows fine control over the visual attributes *via* the Explorer interfaces, although other programs such as XMol could also be used. A further elaboration is to mount more than one data file remotely, as for example with MOPAC coordinates and a MOPAC density matrix, which allows the reader to reconstitute not only the ball-and-stick diagram but also properties such as molecular orbitals and other derived molecular properties.

Chemical Structure Markup Language and 2D NMR Spectral Annotation.—NMR spectroscopy has proved a powerful technique for deriving three dimensional structural information. However, the association of particular features in *e.g.* a two-dimensional spectrum with molecular structure requires substantial imagination. The use of hyperlinks embedded within an image of the spectrum allows a radical departure from the traditional method of presenting such spectroscopic information. A conventional 2D NOESY spectrum *e.g.* of barnase[15] or the oligonuclotide CGCGTTTTCGCG[16] has a small image of the biomolecule inserted into a corner, which in turn is hyperlinked to suitable coordinates (Fig. 3). When the user activates the cursor in this region, the coordinates are downloaded and *via* the mechanism of the *chemical/x-pdb* MIME type, visualized using *e.g.* the RasMol program. This allows the user to select the best viewing angle and representation of the molecule. Selected cross-peaks in the NOESY spectrum, in this case representing inter-chain contacts, are defined as small circular regions on the spectrum corresponding to hyperlinks. This in turn is associated with a file stored on the HTTP server and containing structural markup instructions of the type:

all white wireframe off
ala99 red spacefill on
phe106 red dots on

These indicate the two residues associated with the NOE term, the colour they will be rendered on the user's computer, a representational attribute and its status. This file is allocated a MIME type *chemical/x-csml*, since it formally constitutes what we have defined as a Chemical Structure Markup Language, or CSML. Clearly the present definition must represent only a prototype of such a language and we are indeed currently working on a more formal specification of such markup commands.

When the hyperlink is selected, the csml file is retrieved and used as input data to a small script called csml which is itself resident on the user's local machine and has to be acquired by the user prior to starting the process. The script in turn sends these instructions to the previously activated RasMol process, in this instance using the X-11 inter-client communication protocol implemented with the Tk/Tcl environment, but in principle *via* any process-to-process communication channel. The perception to the user is that as each cross-peak in the window showing the 2D NMR spectrum is selected, the relevant two residues or atoms in the RasMol display window are highlighted virtually instantly. It is important to realize that this process does not occur exclusively on a single computer, but that the information is transferred in realtime using the Internet between two systems that may be continents apart. If the chemical/x-csml mechanism seems excessively complex, this in part is due to the quite stringent requirements for security that the World-Wide-Web protocol mandates. In particular, it is important that only data is transferred transparently between WWW server and client and that no direct instructions are executed on the user's machine without their overt permission and knowledge. Thus although the csml script is executed, the content and action of this script is entirely under the control of the local user and not any remote machine. In future, we anticipate using the recently announced Common-Client-Interface (CCI)[17] to achieve more flexible interaction between the WWW Browser and the molecular viewers.

Within this metaphor, it is easy to imagine the extension to visually complex 3D NMR data, thus enabling a closer association between such data and 3D molecular features. This would of course require the definition of further appropriate secondary chemical MIME types. Currently, NMR data is stored in a variety of proprietary formats and there appears to be urgent need for an open and well defined interchange standard in this area, as for example has been developed for crystallographic data.[18]

Mass Spectral Fragmentation.—Mass spectral fragmentations are also closely associated with structural features in molecules. As the molecule increases in size and complexity, the analysis becomes more difficult. We have demonstrated that the fragmentation pattern of a FAB mass spectrum measured for a permethylated aminocarbohydrate[19] can be hyperlinked to an appropriately annotated 3D structural representation of the compound, in our case again using the RasMol program and the CSML formalism. Since the 3D molecular coordinates are automatically made available to the user, it should be trivial to extend the analysis to performing local computations using this information, as for example in calculating the molecular weight of any particular fragment identified from the hyperlink.

Induced Emission Spectroscopy.—Rotational, vibrational or electronic transitions can be associated with suitable molecular properties using hyperlinks. We have illustrated this with a
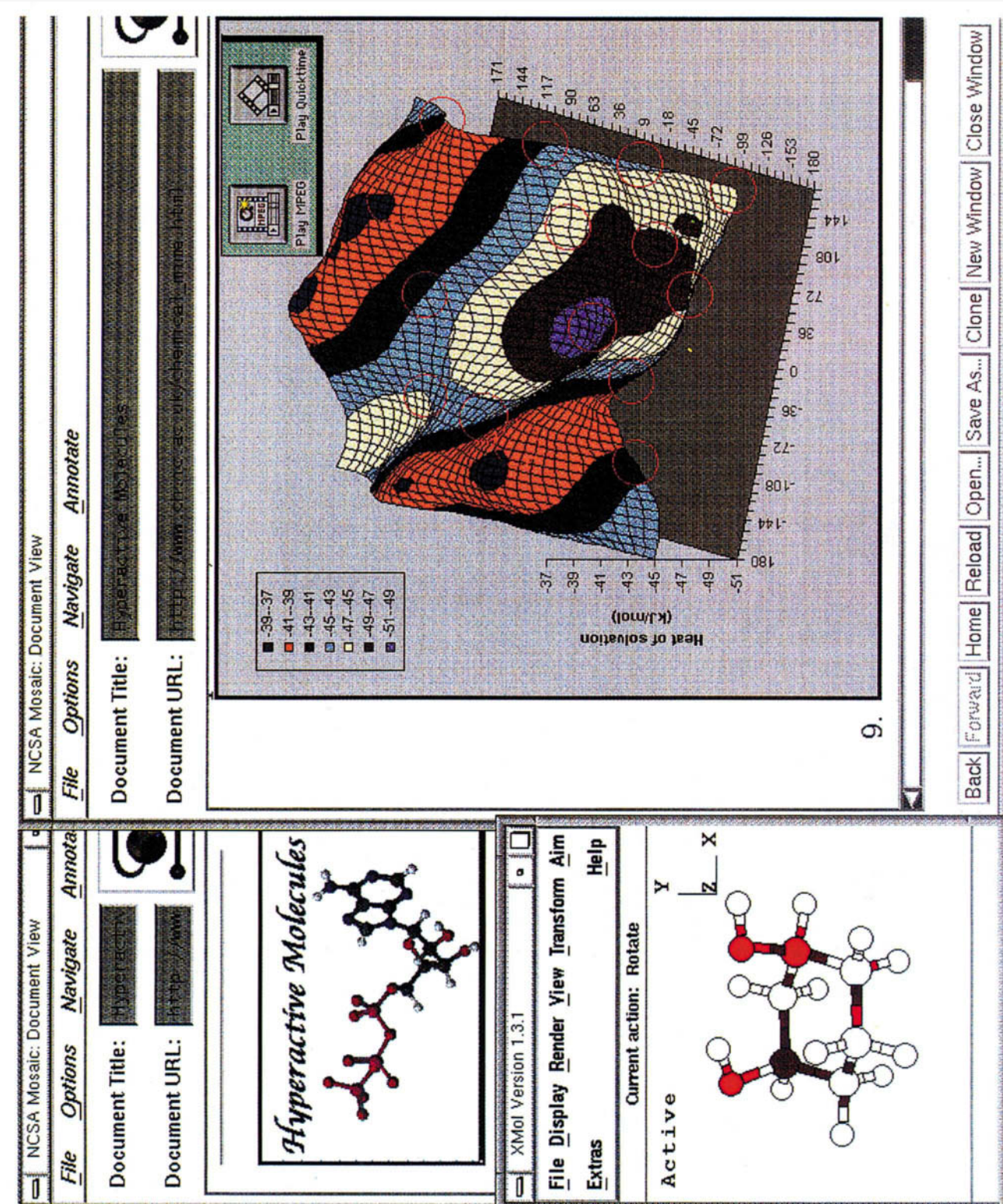
Fig. 2 A PM3/COSMO potential energy surface, with MOPAC coordinates visualized using the XMol program

laser excitation spectrum derived from jet-cooled molecules, for which the rotational fine structure on a vibrational band is evident. The particular rotational transitions can be hyper-linked to the output from a normal coordinate analysis program and which can be animated on the screen to illustrate the relevant normal mode. We have also applied animations to showing the results of molecular dynamics calculations on e.g. the folding of a small oligonucleotide.

Complex Reaction or Metabolic Schemes.—A complex scheme showing e.g. the synthesis of the natural product papuamine[20] can require substantial annotation with reagents, specific com-ments about individual reactions or functional group transform-ations and perhaps reference to crystallographic determinations of key structures and spectroscopic data. This applies equally to complex metabolic schemes, where the substrates and enzymes can be individually annotated. By adding hyperlinks to e.g. a scheme illustrating the citric acid cycle (Fig. 4), such supple-mental data can be viewed when required. More complex hyper-links to remotely stored chemical databases and index based searches, to further such schemes, or to other scientific publi-cations published in hypertext form can be added.

Other Applications of Hyperactive Molecules.—We anticipate that numerous other applications of these techniques might arise. For example, peaks in one and two dimensional chro-matograms could be hyperlinked to appropriate molecular coordinates or amino acid or DNA sequences. The concept that a reader can readily evaluate the conclusions presented in a paper by having access to the original data almost instantly, has profound implications for the quality and reproducibility of chemical information.

Molecular coordinates need not even be stored on any com-puter system, but can be generated as required. For example, the Klotho project[21] may be a forerunner of a system where a chemical name entered by a remote user is automatically translated into a SMILES string using defined grammar and vocabulary and the SMILES string is then used to generate pseudo pdb style molecular coordinates which are then returned to the user in a form suitable for viewing with e.g. RasMol. Such a system could greatly facilitate the semi-automated 'marking-up' of synthetic, spectroscopic and other data.

This system also anticipates how the chemical indexing of molecular data might be achieved using an 'agent' roaming the Internet searching for specified chemical MIME types. Depend-ing on the sophistication of this agent, it is easy to imagine novel correlations between widely disparate molecule data being discovered via such automatic mechanisms. In a sense, the World-Wide-Web becomes another chemistry research tool, from which original and perhaps even general conclusions can be drawn. One might imagine for example such an agent 'discovering' the Woodward–Hoffmann rules by scouring the Internet for all published examples of pericyclic reactions and comparing the three dimensional connectivity data to see if general results regarding the stereochemistry might be inferred.

Molecular Videoconferencing.—A simple extension to the client-server model of information exchange would be to allow peer-to-peer collaborative visualization of molecular data initiated by means of a World-Wide-Web hyperlink. When a hyperlink is activated to display a particular coordinate set, a synchronous window is simultaneously created on one or more remote computer screens associated with collaborators on the project, perhaps with an associated videoconferencing and audio link. The collaborators in effect share a molecular 'white board' within which they can manipulate and edit the molecular data to their joint satisfaction and save the results simul-taneously at both sites. Such methods of working require a

rather more sophisticated network infra-structure, possibly based on ATM (Asynchronous Transport Mode) or IPNG (Internet Protocol New Generation) networking protocols. We have recently demonstrated such a collaborative conferencing session between laboratories in London and Leeds, over the UK's SuperJANET network.

Conclusions
Whilst the mechanisms described in this paper are far from being in common use, the first stage of providing readily accessible 'live' chemical information in a potentially globally searchable form has clearly been achieved. Chemical problems that still need to be solved are how to index such hyperactive molecular information in a chemically cognisant sense and how to improve the metaphors for navigating through such inform-ation such as mapping three dimensional molecular coordinates in a useful sense, e.g. to highlight active sites and other interesting regions. Other issues also require urgent attention. For example, for how long should any given data be expected to reside in automatically accessible form on the Internet? Could it be eventually archived in a manner that would still allow its retrieval using the mechanisms described here. Should the data reside only under the control of 'trusted' agencies, or would data remaining under the control of its originators have to be 'authenticated' using some form of digital signature. Nor is it clear where the cost recovery mechanisms of this new infra-structure would reside. Assuming solutions are found for such problems, it might be anticipated that the conventional scien-tific journal might change more rapidly over the next few years than it has in the last hundred.

Acknowledgements

References
1 F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith and D. G. Watson, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 187.
2 The Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY.
3 H. S. Rzepa, B. J. Whitaker and M. J. Winter, *J. Chem. Soc., Chem. Commun.*, 1994, 1907.
4 For specific details of programs and techniques used, see H. S. Rzepa, *Chemistry and the World-Wide-Web*, in *Chemistry and the Internet*, ed. S. Bachrach, ACS Publications, 1995, to be published.
5 T. J. Berners-Lee, *WWW Names and Addresses, URIs, URLs, URNs*, CERN, Geneva, 1993. See http://info.cern.ch/hypertext/WWW/Addressing/Addressing.html
6 N. Borenstein and N. Freed, Internet Request for Comment (RFC) No. 1521, 1993. See also http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/mailcap.html
7 H. S. Rzepa and P. Murray-Rust, Internet Draft: Chemical Mime Type, May-November 1994. See ftp://cnri.reston.va.us/internet-drafts/draft-rzepa-chemical-mime-type-00.txt
8 R. Sayle (Glaxo Group Research, Greenford, Middlesex, UK). RasMol: A Molecular Visualization System, V2.5, 1994.
9 O. Casher, H. S. Rzepa and S. Green, *J. Mol. Graphics*, 1994, **12**, 226.
10 XMol, Minnesota SuperComputer Center, Minneapolis, MN, USA. See http://www.arc.umn.edu/GVL/Software/xmol/XMol.html
11 B. J. Whitaker, see http://chem.leeds.ac.uk/Project/mopac-form.html
12 J. J. P. Stewart, MOPAC-93. Available from Quantum Chemistry Program Exchange, University of Indiana, Bloomington, USA.
13 D. O'Hagan and H. S. Rzepa, *J. Chem. Soc., Chem. Commun.*, 1994, 2029.
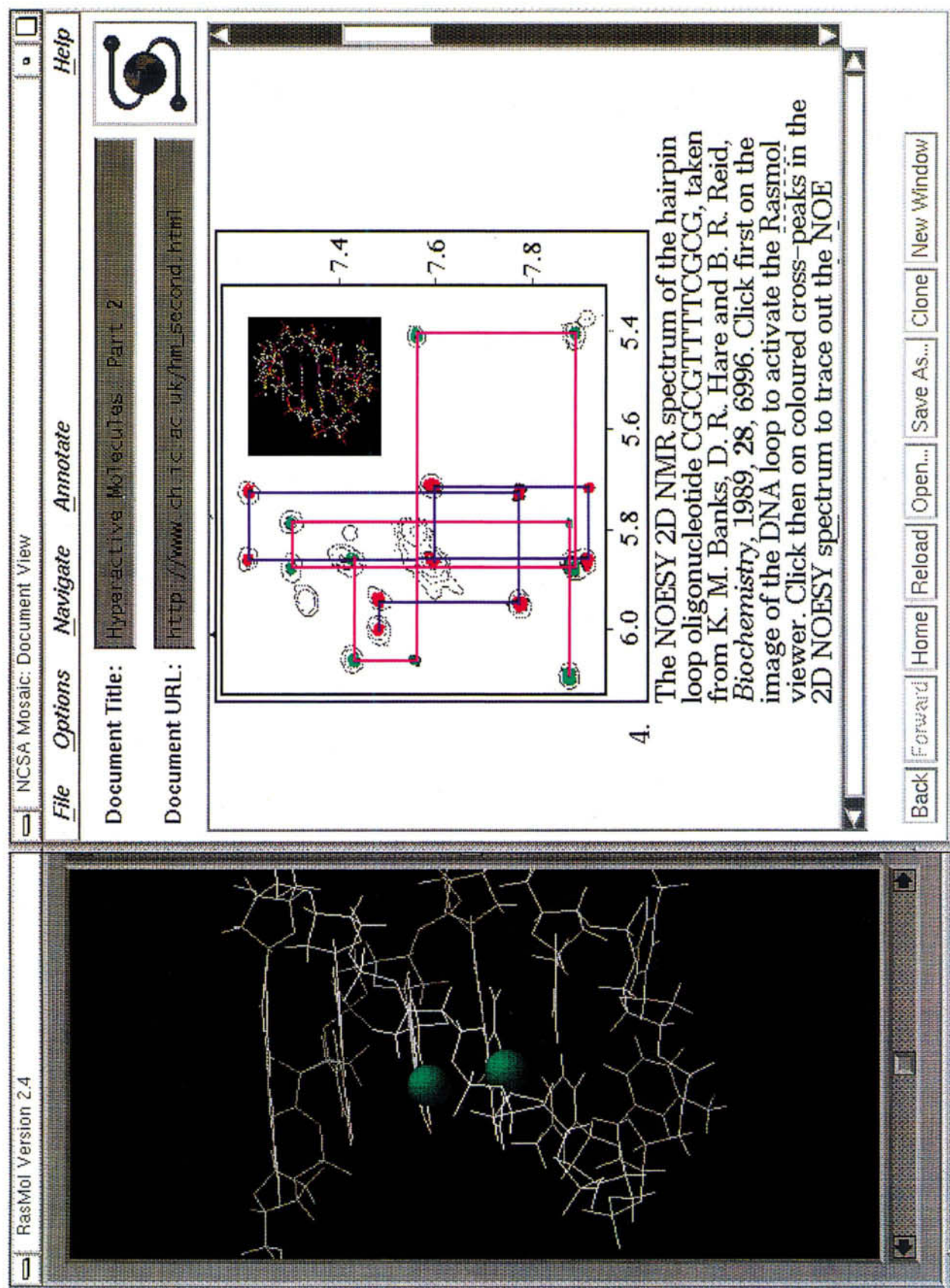14 C. Leach and H. S. Rzepa, First Electronic Computational Chemistry Conference, November, 1994. See http://hackberry.chem.niu.edu/0/ECCC/homepage.html

Fig. 3 The 2D NOESY spectrum of the DNA oligomer CGCGTTTTCGCG, in which the cross peaks are associated with chemical structure markup commands sent to the RasMol viewer

15 M. Bycroft, R. N. Sheppard, F. T.-Kwong and A. R. Fersht, *Biochemistry*, 1990, **29**, 7425.
16 K. M. Banks, D. R. Hare and B. R. Reid, *Biochemistry*, 1989, **28**, 6996.
17 D. Thompson, The CCI Interface, National Center for Supercomputing Applications, University of Illinois, USA. See http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/cci-spec.html
18 S. R. Hall, F. H. Allen and I. D. Brown, *Acta. Crystallogr., Sect. A*, 1991, **47**, 655; F. H. Allen, A. Barnard, A. Cook and S. R. Hall, *J. Chem. Inf. Comput. Sci.*, 1994, submitted for publication.
19 T. Taguchi, A. Seko, K. Kitajima, Y. Muto, S. Inoue, K.-H. Khoo, H. R. Morris, A. Dell and Y. Inoue, *J. Biol. Chem.*, 1994, **269**, 8762.
20 A. G. M. Barrett, M. L. Boys and T. Boehm, *J. Chem. Soc., Chem. Commun.*, 1994, 1881.
21 B. H. Dunford-Shore, F. Fabrizio, J. Holcomb, W. Wise and T. Kazic, Klotho project; Biochemical Compounds Declarative Database, Institute for Biomedical Computing, Washington University; B. H. Dunford-Shore and T. Kazic, Abstract of papers of the American Chemical Society, 1993, **205**, 111. See http://ibc.wustl.edu/klotho/

Fig. 4 An illustration of the citric acid cycle, in which bibliographic information about each substrate and enzyme is collected *via* the activation of a hyperlink in the diagram