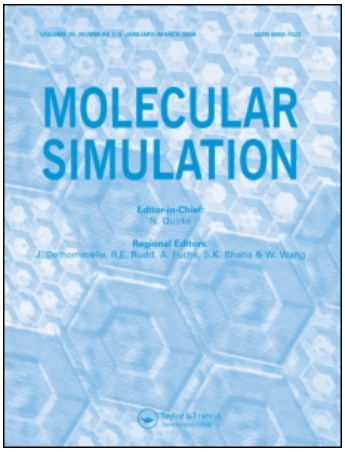


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Computational Chemistry: Application to Biological Systems

Ian R. Gould[a]
[a] Department of Chemistry, Imperial College, London

To cite this Article Gould, Ian R.(2001) 'Computational Chemistry: Application to Biological Systems', Molecular Simulation, 26: 1, 73 — 83
To link to this Article: DOI: 10.1080/08927020108024201
URL: http://dx.doi.org/10.1080/08927020108024201

Molecular Simulation, 2001, Vol. 26, pp. 73–83
Reprints available directly from the publisher
Photocopying permitted by license only

COMPUTATIONAL CHEMISTRY: APPLICATION TO BIOLOGICAL SYSTEMS

IAN R. GOULD*

Department of Chemistry, Imperial College, London SW7 2AY

(Received May 1999; accepted June 1999)

Computational chemistry techniques are discussed in the context of the modelling of biological systems. We consider the application of empirical force fields (Molecular Mechanics) as well as the use of Quantum Mechanical approaches.

Keywords: Molecular Mechanics; Quantum Mechanics; Molecular Dynamics; Proteins; DNA and visualisation

1. INTRODUCTION

The application of computational chemistry to systems of biological relevance encompasses a vast spectrum of techniques attempting to address a large range of issues. These range from trying to understand the relationship between structure and function, including chemical reactivity in proteins, through to the mechanism of protein function and include the application of bioinformatics to enable the searching of large databases. In this article only two subsets of computational chemistry as applied to biological systems will be focussed upon: (1) the application of empirical force fields, Molecular Mechanics (MM) and (2) the application of Quantum Mechanical (QM) methods.

By their very nature biopolymers, proteins and nucleic acids require special techniques to investigate their structure and function, such systems are very large being composed typically of thousands of atoms. Small subtle

*e-mail: i.gould@ic.ac.uk http://www.ch.ic.ac.uk/gould

interaction energies, such as hydrogen bonding and π stacking energies, combine cumulatively to exert large effects on the system. A further compounding issue is that these systems cannot be treated in isolation and require that their environment be incorporated in the simulation, this typically requires inclusion of the solvent either implicitly or explicitly. Typically it is necessary to have a 3-dimensional structure of the system from X-ray crystallography or NMR to enable the simulation of the system under investigation, the absence of such a structure naturally constrains the ability to perform simulations upon it.

In determining what method is most suitable to the investigation of the system in question some very rudimentary decisions need to be made. The primary decision to be addressed is whether the dynamics of the system will be investigated or if it is to be treated as a static system. If the static properties of the system are to be addressed then it is possible to obtain the following information:

- Energies of Interaction, such as the binding of ligands or drug molecules.
- Geometry Optimisation, location of minima.
- Location of Transition States, Activation energies and reaction rates.
- Infra Red (IR), Raman and NMR spectra.

Some of this information can be obtained by both MM and QM methods, whilst some may only be obtained with the latter.

If the dynamic properties of the system are of primary interest then typically this has been the exclusive domain of MM methods, some of the information which can be obtained utilising such methods include:

- Ensemble averages of the structure.
- Folding and unfolding of small biopolymers.
- Time average structures.
- The effects on structure of temperature and denaturants.
- The effects of mutations in the structure of the biopolymer or ligand, this includes the Gibbs free energy changes involved in such mutations.

2. MOLECULAR MECHANICS

Full modelling of proteins and nucleic acids has been for the last 20 years the exclusive domain of Molecular Mechanics due to the large number of atoms involved in the simulations. Typical application of Molecular Mechanics to biological systems has been in the location optimised

geometries, that is the location of local minima, application to obtain time averaged structures and the refinement of X-ray crystal and NMR structures. In order to perform these investigations it is necessary to have an energy functional which relates the energy of the system to the 3-dimensional position of the atoms which compose the system, such a functional is typically described as an empirical force field. An example of an empirical force field being actively used in the investigation of proteins and nucleic acids is the AMBER force field [1] illustrated in Eq. (1)

$$E_{\text{total}} = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{1,3} K_\theta(\theta - \theta_{eq})^2 + \sum_{1,4} V_{n/2}[1 + \text{Cos}(n\varphi - \gamma)] + \sum_{1,4\text{non-bonds}} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{\varepsilon r_{ij}}. \quad (1)$$

The first two terms in Eq. (1) express the bond stretching and bond angle deformation energies as simple harmonic oscillators where the K's represent the force constants for the bond stretch and angle deformation, for a specific pair and triplet of bonded atoms. The third term in Eq. (1) represents the energy associated with "torsional" barriers such as the gauche-trans energy difference in ethane, it is by definition a function of a quartet of atoms. The fourth term in Eq. (1) represents the Van der Waals interaction, between atoms which are not involved in the first three terms, and is therefore described as a 1,4 non-bonded term. The Van der Waals interaction is composed of a long range attractive component and a very short range repulsive component. The final term in Eq. (1) represents the Coulomb, electrostatic, interaction between each atom and every other atom in the system, it is an extremely long range interaction due to the distance between the two atoms occurring in the denominator.

The AMBER force field [1] Eq. (1) is similar to other empirical force fields being used in MM simulations of biopolymers such as CHARMM [2] and MMFF94 [3]. The main difference of "Black Magic" in all such force fields is in the derivation of the underlying parameters that are required. For the bond stretching and angle components it is possible to obtain force constants and equilibrium bond distance and angles from experiment and high level *Ab Initio* Molecular Orbital (MO) calculations. The method for obtaining parameters for the torsional component is less well defined again they may be inferred from experiment or from high level *Ab Initio* MO calculations. The parameters for the Van der Waals component of the force

field can only be obtained from simulations that compare simulated data to experiment. This information enables refinement of the values of the parameters to obtain the best fit to experiment, these simulations are typically performed for neat small organic liquids. The final term in the force field, the Coulomb, is the most difficult to parameterise and is also the most contentious, typically atom centred point charges are calculated by *Ab Initio* MO methods and then transferred to the system of interest.

There are a number of extremely important points to consider when applying empirical force fields to modelling biological systems. First and foremost the very nature of the bond stretching term in Eq. (1) means that bonds cannot be broken or formed in MM simulations, thus chemical reactivity cannot be investigated. There are a large number of parameters which need to be defined for a system, all of the 20 amino acids and 5 nucleic acid units are usually predefined for a particular force field. Investigation of a system that includes residues not predefined requires a large amount of expertise to parameterise the system. Parameterisation is usually performed for a particular conformation, environment or set of conditions, therefore, attempting to perform simulations in which the system is far removed from the prevailing conditions of parameterisation may lead to very erroneous results.

Since for MM the energy is a functional with respect to the nuclear co-ordinates then it is a relatively simple process to take the first derivative to obtain the forces on the atoms and to proceed to minimise the structure to a obtain a local minima. By analogy the second derivative at a stationary point indicates whether the structure is a true minima or a saddle point, the resulting Hessian gives the force constants and hence the Infra Red spectra of the molecule under investigation. The relative mathematical simplicity of the force field equates to simple algorithmic implementation of the energy, first and second derivatives in a computer package. In addition the low number of floating point operations required for these tasks facilitates their very rapid evaluation on the majority of today's workstations and parallel computers, thus enabling the study of systems composed of a very large number of atoms. This enables the investigation of the time evolution of such systems through the application of Molecular Dynamics (MD).

3. MOLECULAR DYNAMICS

The primary function of Molecular Dynamics (MD) is to enable the investigation of the time evolution of a system, obtaining the atomic

positions and velocities as a function of time. To perform MD simulations it is necessary to integrate Newton's equations of motion Eq. (2)

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i [U(r_1, r_2, \ldots, r_n], i = 1, n \tag{2}$$

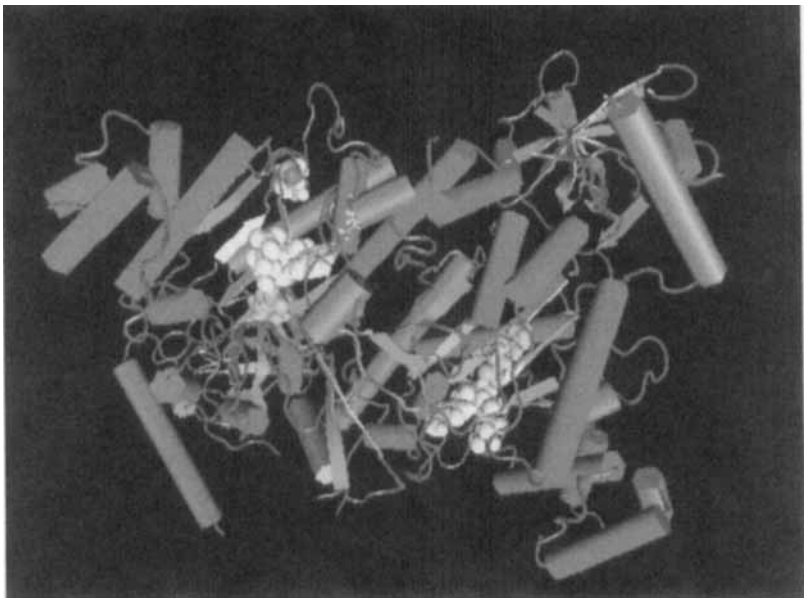
where the right hand-side of (2) are the gradient or forces on the atoms. The integration is performed numerically and this limits the time-step that may be used for integration, a value of 1 femtosecond (fs), 1×10^{-15} s, is usually used. The combination of a very small time-step and the large size of the system under investigation puts constraints on the size of system that is amenable to simulation, for systems of ca 20000–30000 atoms a realistic simulation length of a couple of hundred picoseconds is fairly normal.

The typical protocol for simulating a biopolymer, protein or nucleic acid, is as follows

- Obtain an initial structure from X-ray crystal or NMR data.
- Immerse the biopolymer in a pre-equilibrated box of solvent molecules, typically water.
- Perform limited geometry optimisation of complete system.
- Perform MD until the system is equilibrated to the desired temperature and density.
- Perform MD with data collection of positions, velocities, densities, pressures and temperature *etc.*
- Data analysis to obtain ensemble averages of desired properties.

One of the most important applications of MD to biopolymers is in the prediction of the effect of a mutation on a known native structure. Due to advances in protein engineering it is now possible to selectively substitute different amino acids into a protein sequence, a process known as site directed mutagenesis. By performing MD simulations on both the native structure and the mutant it is possible to interpret the effect of the mutation on the catalytic efficiency of the protein by analysing the effect on the 3-dimensional structure.

The results of MD simulations are illustrated by two movies associated with the article which are available from the author. The movie LYSU.mov shows the first 100 ps of an all atom simulation of the enzyme LysyIRS this is a large dimer protein which is responsible for producing 4P4A an important factor in platlet aggregation, stress response and cell proliferation. A cartoon of the dimer is shown in Figure 1. The protocol for the simulation



was as follows:

- Dimer structure obtained from Brookhaven data base.
- Dimer immersed in a box of Monte Carlo equilibrated waters.
- System size ~100,000 atoms of which 29,000 are protein the remainder water.
- System subjected to a few hundred cycles of geometry optimisation using steepest descent algorithm.
- System equilibrated until constant temperature and density achieved.
- 100 ps "production", data collection, performed.
- 1 fs time-step used, 14 Angstrom non-bonded cutoff, AMBER all atom force field.

Due to the extreme size of the system only the protein has been visualised in the movie.

To illustrate the inclusion of solvent in MD simulations we have included a second MD movie of Myoglobin, MYO1.mov. Figure 2 illustrates the structure of the protein and surrounding solvent. The protocol for this simulation was identical to the LYSU protocol above, the only difference being the number of atoms in the simulation.

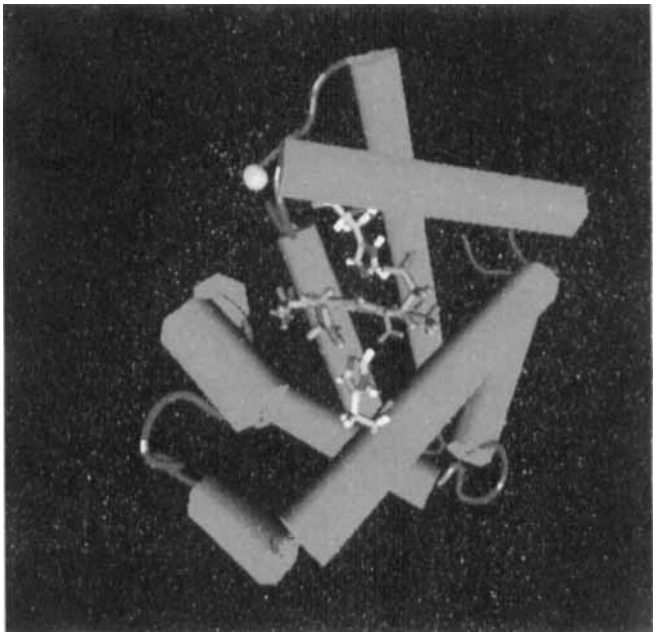


FIGURE 2 Snapshot of Myoglobin simulation. (See Color Plate XVI).

4. QUANTUM MECHANICS

It has been shown in the last section that Molecular Mechanics is incapable of accounting for chemical reactivity in biological systems due to the fact that it does not explicitly include terms involving the dynamic distribution of electrons. To address the "Electronic" problem, that is a method that can be applied to electrons, one has to apply a Quantum Mechanical approach. The starting point of all QM method is Schrödinger's wave Eq. (3)

$$\hat{H}\Psi = E\Psi. \tag{3}$$

The method of solution of this equation is broadly speaking split into two different approaches, *Ab Initio* and Semi-empirical Molecular Orbital (MO) techniques. Common to both is the ability to reproduce experimental energetics, geometries and spectra (Infra Red, Ultra Violet, *etc.*), in principle any property of interest may be obtained. In particular the main feature of both approaches is that reactions can be modelled. The

fundamental difference between *Ab Initio* and Semi-empirical is that the former relies on no "empirical" parameters whilst the latter uses some experimentally determined data. A further benefit of using these methods is that excited states may be investigated, this is important as excited states are important in processes such as electron transfer in photosynthetic systems.

There are a large number of methods available for performing *Ab Initio* electronic structure calculations these include

- HF-SCF – Hartree-Fock Self-Consistent-Field the simplest level of theory available.
- DFT – Density Functional Theory – the lowest cost method to include electron correlation.
- MP2, MP3, MP4 – Moller-Plesset perturbation theory, derivative of Rayleight-Schrödinger perturbation theory, good for accounting for dynamic correlation.
- CASSCF-Complete Active Space Self Consistent Field theory, a good method for investigating non-dynamic correlation, chemical reactivity.

The major problem with all of these methods is that the computational complexity scales very poorly with increased system size. This is due to the fact that a set of basis functions have to be used to represent the atomic orbitals of the atoms in the system, for N-basis functions the cheapest form of *Ab Initio* MO theory HF-SCF scales at best as $N^{2.7}$. This has typically constrained the number of atoms in the system to be less than 100 pre-1982. However, the development and application of the so-called direct methods found in such packages as Gaussian 98 [4] and Q-CHEM [5] has made a significant difference to the size of system to which the methods can be applied, systems up to ca 500 atoms. Further recent developments of the Fast Multipole Method and methods for efficient evaluation of the two-electron integrals, the major computational bottleneck, is pushing this limit into the low thousands of atoms.

Very briefly here we describe an application of *Ab Initio* MO techniques to the investigation of a biological system. The application is the investigation of the excited states of the Iron containing Heme prophyrin of Myolgobin. Using the crystal structure of Horse Heart Myoglobin the excited states of the Iron heme unit have been calculated using the Configuration Interaction Singles (CIS) MO method. Comparison with experimental determined energy gaps and oscillator strengths are given in Table I. Figures 3 and 4 illustrate the HOMO and LUMO for the

TABLE I Comparison of experiment and theory

Band property (e.V.)	*Calculated*	*Experiment*
QY-QX	0.07	0.12
Oscillator (QY/QX)	2.36	2.60
Soret (X-Y)	0.04	021
Osc Soret (Y/X)	1.98	1.89
Soret (Y-QY)	2.47	0.65
Oscillator (Soret Y/QY)	21	20–25

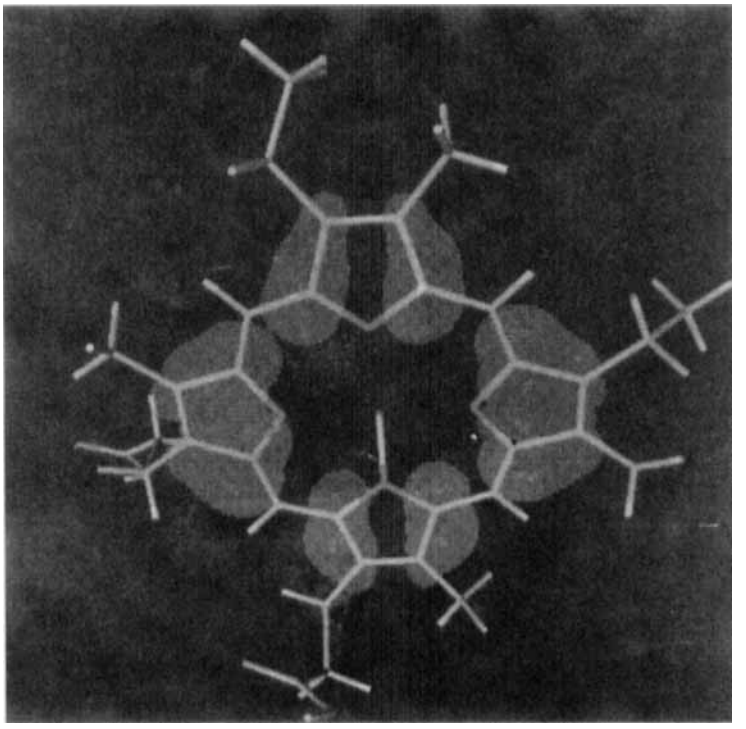


FIGURE 3 HOMO Orbital of Heme unit of Myoglobin. (See Color Plate XVII).

system, interpretation based on these orbitals is instructive in understanding the excited states.

5. NEW DEVELOPMENTS AND THE WAY FORWARD

One of the most interesting developments for the application of computational chemistry to biological systems is the development of hybrid QM/

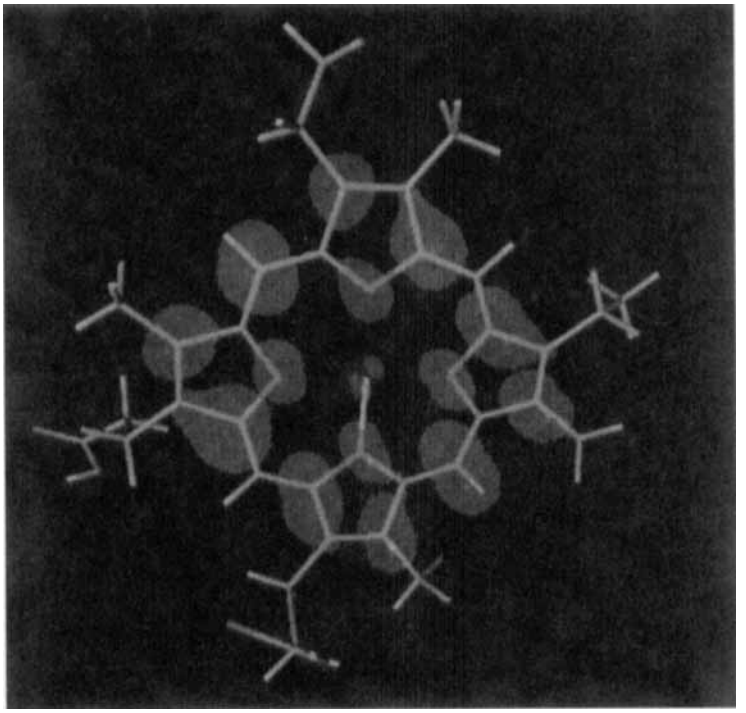


FIGURE 4 LUMO orbital of Heme unit of Myoglobin. (See Color Plate XVIII).

MM methods. The approach is a divide and conquer one where the interesting electronic component of the system, *i.e.*, the area in which the reaction is occurring, is treated with a MO method whilst it is influenced by a MM description of the rest of the system. In addition to this type of approach, recent developments in Semi-empirical methods are enabling the investigation of systems composed of ca 10,000 atoms, similar techniques are being developed for *Ab Initio* MO methods, in particular the Fast Multipole Method (FMM), which ultimately will enable similar size calculations.

Inherent to these advances is the year on year increases in the speed of computers, particularly with the development of high performance parallel machines, which again is increasing the size of system amenable to calculation. This requirement for high performance computing resources has recently been demonstrated by Duan and Kollman [6] who have performed a classical MD simulation on a small protein for a microsecond, the simulation consisting of $\sim$20,000 atoms, to try to access the process of

protein folding. Such a simulation would be impossible without massively parallel supercomputer technology.

Acknowledgements

The simulation of Lys-U is the work of Samantha Hughes. The Myoglobin simulation and electronic structure calculations is the work of Halima Amer.

References

[1] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. (1995). "2nd Generation force-field for the simulation proteins, nucleic-acids, and organic-molecules", *J. Am. Chem. Soc.*, **117**, 5179.

[2] Mackerell, A. D. (1998). "Developments in the CHARMM all-atom empirical energy function for biological molecules", *Abstracts of papers of the American Chemical Society*, **216**, 42.

[3] Halgren, T. A. (1999). "MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries", *J. Comp. Chem.*, **20**, 730.

[4] Q-chem 1.1, A Quantum Leap into the Future of Chemistry, Johnson, B. G., Gill, P. M. W., Head-Gordon, M., White, C. A., Baker, J., Maurice, D. R., Adams, T. R., Kong, J., Challacombe, M., Schwegler, E., Oumi, M., Ochsenfeld, C., Ishikawa, N., Florian, J., Adamson, R. D., Dombroski, J. P., Graham, R. L. and Warshel, A., Q-Chem, Version 1.1, Q-Chem, Inc., Pittsburgh, PA (1997).

[5] Gaussian 98, Revision A.5, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Zakrzewski, V. G., Montgometry, J. A., Jr. Stratmann, R. E., Burant, J. C., Dapprich, S., Millam, J. M., Daniels, A. D., Kudin, K. N., Strain, M. C., Faraks, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., CLifford, S., Ochterski, J., Petersson, G. A., Ayala, P. Y., Cui, Q., Morokuma, K., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Cioslowski, J., Ortiz, J. V., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, R., Gomperts, R., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Andres, J. L., Gonzalez, C., Head-Gordon, M., Replogle, E. S. and Pople, J. A., Gaussian, Inc., Pittsburgh PA, 1998.

[6] Duan, Y., Wang, L. and Kollman, P. A. (1998). "The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation", *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 9897.