Markov Chain Monte Carlo for Internal Dosimetry on a Supercomputer Cluster

CHAPTER · JANUARY 2005

DOI: 10.1021/bk-2007-0945.ch007

4 AUTHORS, INCLUDING:

Guthrie Miller
Jubilado, Santa Fe, United States
116 PUBLICATIONS **1,004** CITATIONS
SEE PROFILE

Raymond Guilmette
Lovelace Respiratory Research Institute
185 PUBLICATIONS **1,587** CITATIONS
SEE PROFILE

ACS SYMPOSIUM SERIES **945**

Applied Modeling and Computations in Nuclear Science

Thomas M. Semkow, Editor
New York State Department of Health

Stefaan Pommé, Editor
Institute for Reference Materials and Measurements

Simon Jerome, Editor
National Physical Laboratory

Daniel J. Strom, Editor
Pacific Northwest National Laboratory

Sponsored by the
ACS Divisions of Nuclear Chemistry and Technology and Computers in Chemistry

American Chemical Society, Washington, DC

Chapter 7

Markov Chain Monte Carlo for Internal Dosimetry on a Supercomputer Cluster

G. Miller, L. Bertelli, T. Little, and R. Guilmette

Los Alamos National Laboratory, MS–E546, Los Alamos, NM 87545

The methods used at Los Alamos for calculating internal dose for plutonium and americium are described. The main method, the ID code, is a straightforward use of Bayes' theorem, evaluated using Markov Chain Monte Carlo. A supercomputer cluster is used to do mass calculations on many cases at once. As an alternative, a single workstation continually does calculations as new bioassay data comes in, spreading out the calculation load over the year. Both methods are being used successfully.

Introduction

Internal dosimetry is concerned with the problem of determining the radiation dose to workers caused by forms of radiation that cannot be measured directly (as with a dosimetry badge). If, for example, the α-emitting nuclide ^{239}Pu is inhaled, it will impart radiation dose to the lungs, and after dissolving will be absorbed to blood and deposited in the bone and liver, imparting dose to these organs. Monitoring for exposure to ^{239}Pu is done by measuring the ^{239}Pu content in various kinds of samples (bioassay), for example, urine, fecal, lung count, etc.

The measurements are interpreted using biokinetic models that describe how ^{239}Pu is transported through the body. The biokinetic models describe how a unit

amount of material taken into the body in a certain way (for example, inhalation) will later in time appear in various bioassay compartments (for example, the lungs, urinary excretion) and how radiation dose will be accumulated in the course of time in the different body organs and tissues.

Standard biokinetic models have been proposed by the International Commission on Radiation Protection (ICRP) (*1,2*).

Given a set of agreed-upon biokinetic models, the inverse problem of internal dosimetry is to use the bioassay measurements to infer if and when intakes may have occurred and the magnitude of the resultant radiation dose to the worker. Intake-based biokinetic models are used to determine the time and amount of intakes and to calculate the 50-year effective whole body dose to the worker (the CEDE, Committed Effective Dose Equivalent) associated with each intake. The process obviously entails considerable uncertainty, so quantitatively assessing uncertainty is also of great importance. For the past decade, we have been pursuing a Bayesian statistical approach to this problem. Other approaches to internal dosimetry involve an expert assessor evaluating each case, as described, for example, in Ref. (*3*).

Bayesian Approach to Internal Dosimetry (ID code)

In the problem of internal dosimetry there are M bioassay data y_j taken at times t_j for $j = 1,\ldots, M$. These data are used to determine N possible intakes with amounts ξ_i, biokinetic types l_i, times of intake t_i, for $i = 1,\ldots, N$. The intake times are ordered, so that $t_1 < t_2 < \ldots < t_N$. The domain of time t_i is the time interval Δt_i. That is, t_i is in the interval Δt_i. The intervals Δt_i cover the time domain of all possible intakes in a nonoverlapping and ordered way. The time intervals are usually chosen to be the times between successive bioassay measurements, in which case $N = M - 1$. The time intervals are chosen to be sufficiently small so that the probability of multiple intakes may be neglected in any interval.

Using the notation:

$$Y \equiv \{y_1, y_2, \ldots, y_M\}$$
$$\Xi \equiv \{\xi_1, l_1, t_1, \ldots, \xi_N, l_N, t_N\},$$

the problem is to determine the parameters Ξ from the data Y.

Using Bayes theorem, the probability distribution of Ξ given Y can be immediately written down as

$$P(\Xi \mid Y) \propto P(Y \mid \Xi) P(\Xi), \tag{1}$$

that is, the probability of particular values of the parameters given the data is proportional to the probability of the measured values of the data given the parameters (the likelihood function) times the prior probability of the parameter values.

The calculational problem is then to integrate over the full detailed posterior probability distribution function, given by Eq. 1, in order to determine the marginal probability distribution of quantities of interest (for example, some dose quantity of interest). This multi-dimensional integration problem can be solved with the Markov Chain Monte Carlo Method (*4*) using the Metropolis algorithm (*5*). A Markov chain is a sequence of random variables Ξ_k such that Ξ_k depends on its predecessor Ξ_{k-1} and does not depend further on the history of the chain.

The likelihood function $P(Y|\Xi)$ is of the form

$$P(Y \mid \Xi) \propto \exp\left(\sum_{j=1}^{M} L_j(\Xi)\right),$$

because of the assumed independence of the M measurements, where $L_j(\Xi)$ is the log-likelihood function for the j^{th} measurement. Exact numerical calculation of the Poisson or Gaussian/log normal likelihood function is used (*6*).

The prior probability distribution $P(\Xi)$ is taken to be of the form

$$P(\Xi)d\Xi = \prod_{i=1}^{N} P(\xi_i)d\xi_i P(l_i)P(t_i)dt_i\,,$$

assuming independent probabilities for intake amount, biokinetic type, and time of intake for each intake.

The prior probability distribution of biokinetic types l is a discrete probability distribution over $\{l_1, l_2, \ldots, l_{ni}\}$, usually uniform except that the ICRP-recommended default model is given a higher probability.

The prior probability distribution for ξ_i and t_i depends on whether or not a known incident has occurred in the intake time interval Δt_i. This and other details are discussed elsewhere (*7*).

In the Markov Chain Monte Carlo method, values of the parameters Ξ are moved step-wise around in their multidimensional phase space (each chain iteration is one step) in such a way that the probability of Ξ being in a some region of phase space is proportional to the posterior probability given by Eq. 1. In effect, parameter values are randomly generated from the posterior distribution. Thus, to examine the distribution of some quantity of interest, the quantity is merely calculated for each step of the chain, and a histogram made of the results. Every quantity has a distribution, and this distribution is a direct representation of the uncertainty of the quantity.

The distributions obtained as described above depend, in principle, on the number of chain iterations, the chain starting point and the random number generator seed. The chain needs to be run long enough so these dependencies are minimal. To determine chain convergence, two separate calculations are done, with widely disparate starting points and different random number generator seeds. The results of the two calculations must be sufficiently close for the

calculation result to be considered to be "converged". If this is not the case, the chain is run for more iterations.

The Lambda Cluster at Los Alamos

There are a number of computer clusters in use at Los Alamos. Lambda is an unclassified general computing resource. It consists of 328 Intel Pentium 3 processors across 164 compute servers (Compaq DL360s) running Redhat Linux operating system. The processors are 1 Ghz each and every node has 1 GB of memory. Lambda uses a Gigabit Ethernet interconnect.

Lambda uses software called LSF (Load Sharing Facility) version 4.1 to control user jobs, which are submitted using the BSUB command. This software is written by Platform Computing (platform.com). All jobs are submitted to the cluster through queues which provide access to various machine groups, user groups, or resources. LSF schedules and runs jobs, distributing them across the compute nodes using features such as queue or machine limits, queue priorities, processor reservation, and job backfilling to provide efficient utilization of the cluster.

There are three main queues on Lambda, the interactive queue, the large queue (maximum job time 2 days), and the long queue (maximum job time 7 days). The ID code calculations are run in the large queue. Under optimal conditions, about 100 processors are available. A small fraction of the jobs require more than 2 days to complete (most require much less). The long running cases involve hundreds of bioassay data points spanning up to 50 years of work history. The jobs that fail to complete in 2 days are restarted in the long queue, where at most about 30 processors are available.

It has been found that cluster computing is not 100 % reliable, and methods for checking that jobs have completed satisfactorily have been developed. The small fractions of cases with problems are identified, and the calculations rerun.

Using Lambda under optimal conditions, the plutonium and americium internal dosimetry calculations for all employees who have submitted bioassay samples in the past year (some 2000 persons) can be completed in 2 days, with a small handful of cases requiring about a week. Thus, the entire task can be accomplished within one week.

Comparisons with Single Workstation Calculations

The other approach to ID code calculations used at Los Alamos is to use a single desktop workstation that runs calculations whenever new bioassay data become available. A dual 3.6 Ghz processor with hyperthreading is used so that

there are 4 effective processors, two of which are used for the ID code (convergence testing requires two runs), and two are available for other purposes. Direct speed comparisons are shown in Table I.

Table I. Processor Speeds Determined From Wall-Clock Comparisons

Processor	*Relative processor speed*
Lambda	1
3.6 Ghz Intel Xeon	2.1
3.6 Ghz Intel Xeon Hyperthread	1.2

In the single-workstation mode of operation, a single dual-processor workstation continually does calculations as new data come in, spreading out the calculational load over the year. This workstation is set up with hyperthreading so it effectively has 4 processors, and therefore allows other work to be done as well. The advantage is a simpler, privately owned computing environment. The disadvantage is that it would be impossible to recalculate all cases in a short time should this be necessary.

A comparison of local calculations with supercomputer calculations is shown in Table II.

Table II. Correlation Coefficient Between Workstation And Supercomputer ID Code Results

Quantity	*Number of cases*	*Correlation coefficient*
CEDE	22079	0.9997
total CEDE	1967	0.9998

The correlation coefficient (see, for example, matheworld.com) is 1 when the two results are exactly equal. Table II shows a comparison for 1967 person, nuclide combinations. The first row shows the year by comparison of calculated CEDE for each of these person-nuclide combinations over all the years covered by the bioassay data. The reason there is any disagreement at all is that the two calculations effectively involve different random number seeds. This is because the Markov Chain Monte Carlo calculation is preceded by another Monte Carlo calculation-of the exact likelihood functions. Since these calculations are usually not done in exactly the same sequence, the random number sequence for the Markov Chain calculation is different.

Comparisons with Faster, Approximate Method (UF code)

Another code, the UF or unfolding code, uses an approximate Bayesian method (*8*). The basic calculation is of a single intake, with the bioassay tail from previous intakes subtracted. An iteration process is used to eliminate intakes with large probability of the hypothesis "no intake" or "flatline" interpretation of the data relative to the hypothesis that an intake has occurred as calculated. The UF code uses numerical integration and, because of the necessity of using Gaussian uncertainty propagation in the tail subtraction, does not use the exact likelihood calculation.

The major flaw of the UF code is in the assignment of the times of intakes. The differences between the two codes are illustrated by using the test dataset shown in Fig. 1.

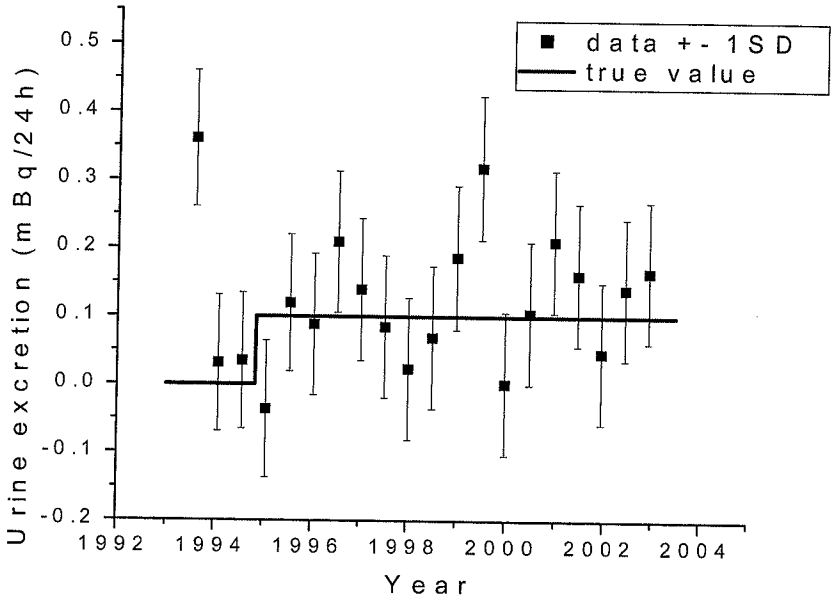

Figure 1. Numerically generated test dataset.

The data are generated from a Gaussian distribution with standard deviation 0.1 mBq/24 hr. After the 3rd data point, the data are shifted upwards by 0.1 mBq/24 hr.

The dose conversion factor from urine excretion a long time after the intake to E(50) (same as CEDE or 50-year whole body effective dose, except using a particular set of tissue weighting factors defined in ICRP publication 60) is about 50 Sv per Bq/24 hr for type S, 5 micron ^{239}Pu, so a urine excretion of 0.1 mBq/24 hr corresponds to a 5 mSv intake.

The ID code result for yearly CEDE (50-year effective, whole-body dose) is shown in Fig. 2. As for all quantities, the yearly CEDE's shown in Fig. 2 are obtained from distributions. The grey bars show the 5 % and 95 % probability limits of these distributions, while the dark squares shows the average values. Note that the lower probability limit is very small in all cases, meaning that there is not certainty that an intake occurred in any particular year. In 1997 the average value exceeds the 95 % limit. This is possible for a distribution with greater than 95 % probability concentrated near zero and a small tail. The time of intake is very uncertain.

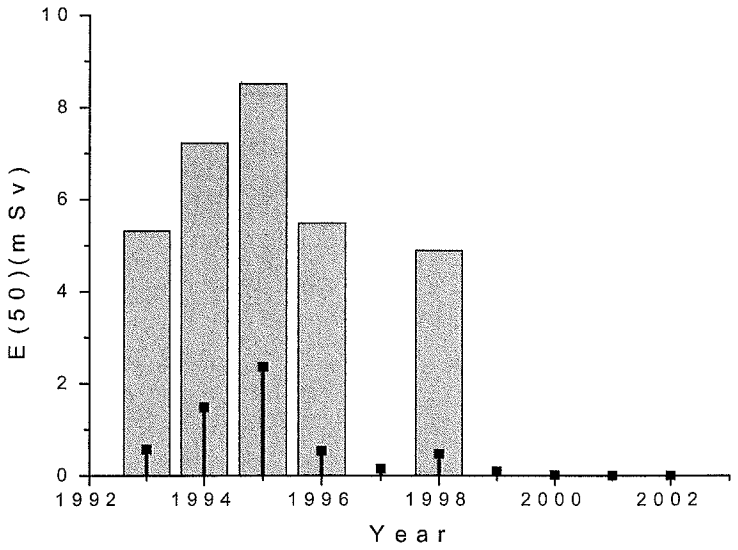

Figure 2. Yearly CEDE's calculated by ID code for test dataset.

The calculations assume an "alpha prior" for intake amount ξ

$$P(\xi) = \frac{\alpha \Delta t}{\xi} \left(\frac{\xi}{A} \right)^{\alpha \Delta t},$$

where Δt is the time interval in which the intakes occur and A is an unimportant normalizing constant, with the probability of intake per year $\alpha = 0.001$/yr (*9*), so that the probability of more than one significant intake is small (even though 19 intakes are allowed in the calculation). The distributions of CEDE for different

years shown in Fig. 2 are anticorrelated, so that they represent the uncertainty in the time of intake of a single intake rather than uncertainties of multiple intakes.

The prior probability distribution of time of intake is assumed to be uniform in each interval. The biokinetic type prior consists of the 6 biokinetic types for inhalation: dissolution type M and S, and particle size 1, 5, and 10 μm AMAD (Activity Median Aerodynamic Diameter).

The cumulative probability for the total CEDE from all intakes is shown in Fig. 3. As seen from Fig. 3, the total CEDE is calculated to be in the range from 1 to 10 mSv, with a central value near the true value of 5 mSv.

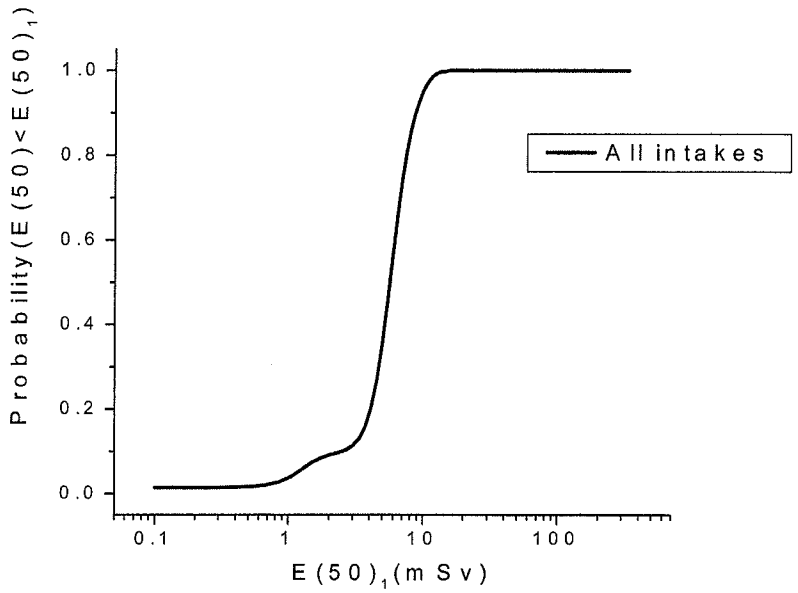

Figure 3. Cumulative distribution of total CEDE, from ID code calculation for test dataset.

The UF code result for this case is shown in Fig. 4. The curve shows the calculated average excretion. The calculated average CEDE is 2 mSv, with credible limits from 0 to 9 mSv. The main error is in the calculated time of intake, and there is no indication of the uncertainty of time of intake using the UF code.

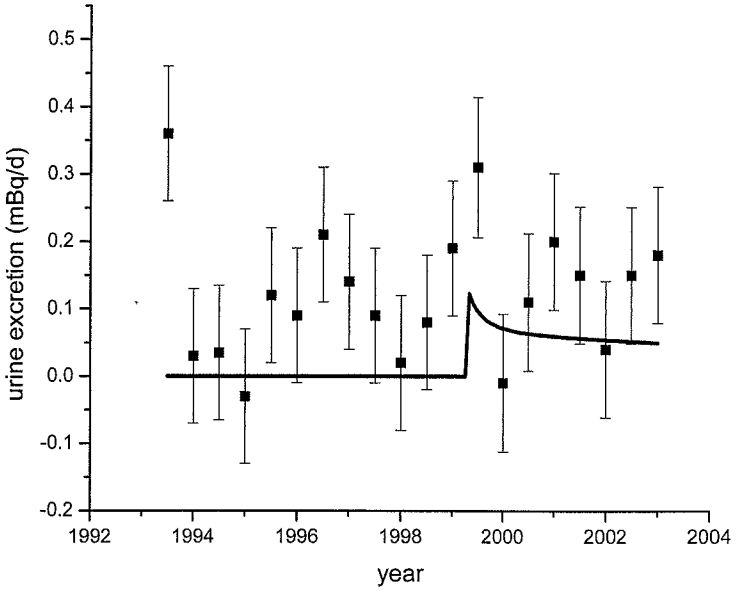

Figure 4. UF code calculation for test dataset.

Nonetheless, the correlation between ID and UF code results is rather good. Table III shows the correlation coefficients for a number of comparisons. Note that the annual dose (for the year of the last bioassay sample) and the total CEDE for all intakes are in very good agreement.

Table III. Correlation Coefficient For ID Code (Lambda) And UF Code Results

Quantity	*Number of cases*	*Correlation coefficient*
CEDE	40777	0.6964
total CEDE	3856	0.9981
annual dose	3856	0.9859

Figure 5 shows a plot of total CEDE from the UF code versus that from the ID code. The year-by-year CEDE is less well correlated. This can be understood from the test case shown in Figs. 1 and 2, where the average CEDE from the ID code is quite different from the UF code result. This agreement is really quite remarkable in that the numerical methods are completely different, and the UF code does not use the exact likelihood calculation.

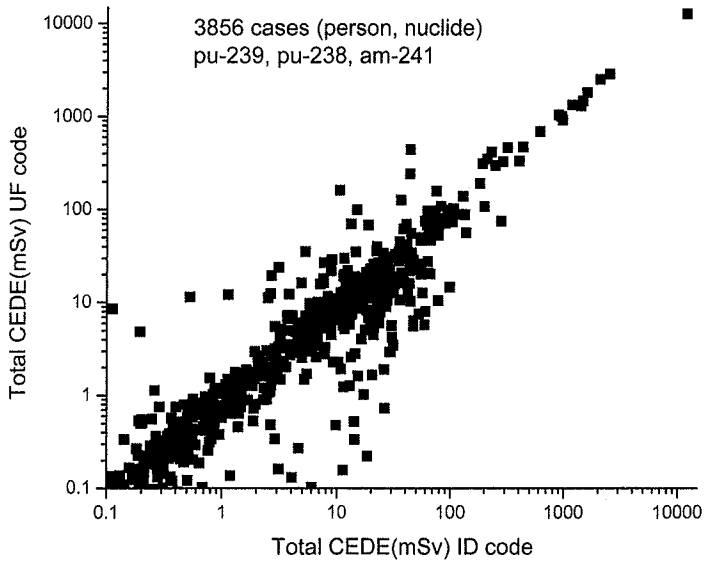

Figure 5. Comparison of total CEDE from ID and UF codes.

Discussion

The methods in use at Los Alamos for calculation of internal dose for plutonium and americium have been described. Statistical inference based on Bayes' theorem is used. Some advantages of this approach are the following:

1. It is scientifically defensible; indeed, it is definitive.
2. Professional judgment and prior knowledge are not mixed up with statistical science; they are confined to defining or specifying the prior probability distributions.
3. The full information content of the bioassay measurements is utilized; the exact likelihood calculations focus attention on the measurement science.
4. The uncertainty of yearly CEDE is often very large; this uncertainty is directly calculated.
5. Inexpensive batch processing of large numbers of cases is possible.

Straightforward parallel processing, one case to each processor, is used on a supercomputer cluster for mass calculations of large numbers of cases. As an alternative, a single workstation continually does calculations as new bioassay

data comes in, spreading out the calculational load over the year. Both methods are being used successfully.

Comparisons between the ID and UF codes show excellent agreement and serve as a validation of both codes, since completely different numerical methods are used in each code.

The yearly CEDE often has a very large uncertainty. This is an unavoidable fact. It is perhaps a problem that current regulations are based on yearly CEDE rather than, say, total CEDE or annual dose, which was used before 1992. The agreement between the ID and UF codes shows that total CEDE and annual dose are easier to calculate using approximate methods; furthermore, they have much smaller uncertainties.

References

1. *Individual Monitoring for Intakes of Radionuclides by Workers: Design and Interpretation*; ICRP Publication 54; Vol. 19 of Annals of the International Commission on Radiation Protection; Pergamon Press: Oxford, 1988.
2. *Individual Monitoring for Intakes of Radionuclides by Workers: Replacement of ICRP Publication 54*; ICRP Publication 78; Vol. 27 of Annals of the International Commission on Radiation Protection; Pergamon Press: Oxford, 1998.
3. Marsh, J. W.; Baily, M. R.; Birchall, A. A step-by-step procedure to aid the assessment of intake and doses from measurement data. *Radiation Protection Dosimetry* **2005**, *114*, 491-508.
4. Gilks, W. R.; Richardson, S.; Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: New York, NY, 1996.
5. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; A. H. Teller; E. Teller. Equation of state calculation by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.
6. Miller, G.; Martz, H. F.; Little, T. T.; Guilmette, R. Using exact Poisson likelihood functions in Bayesian interpretation of counting measurements. *Health Phys.* **2002**, *83*, 512-518.
7. Miller, G.; Martz, H. F.; Little, T. T.; Guilmette, R. Bayesian internal dosimetry calculations using Markov chain Monte Carlo. *Rad. Prot. Dosimetry* **2002**, *98*, 191-198.
8. Miller, G.; Inkret, W. C.; Martz, H. F. Internal dosimetry intake estimation using Bayesian methods. *Rad. Prot. Dosimetry* **1999**, *82*, 5-17.
9. Miller, G.; Inkret, W. C.; Little, T. T.; Martz, H. F.; Schillaci, M. E. Bayesian prior probability distribution for internal dosimetry. *Rad. Prot. Dosimetry* **2001**, *94*, 347-352.

Reprinted from ACS Symposium Series 945
Applied Modeling and Computations in Nuclear Science
Thomas M. Semkow, Stefaan Pommé, Simon M. Jerome,
and Daniel J. Strom, Editors
Published 2006 by the American Chemical Society