Copyright © 2002 Tech Science Press CMES, vol.3, no.5, pp.557-574, 2002

JavaGenes: Evolving Molecular Force Field Parameters with Genetic Algorithm

Al Globus[1], **Madhu Menon**[2], **and Deepak Srivastava**[1]

Abstract: A genetic algorithm procedure has been developed for fitting parameters for many-body interatomic force field functions. Given a physics or chemistry based analytic form for the force field function, parameters are typically chosen to fit a range of structural and physical properties given either by experiments and/or by higher accuracy tight-binding or ab-initio simulations. The method involves using both near equilibrium and far from equilibrium configurations in the fitting procedure, and is unlikely to be trapped in local minima in the complex many-dimensional parameter space. As a proof of concept, we demonstrate the procedure for Stillinger-Weber (S-W) potential by (a) reproducing the published parameters for Si by using S-W energetics in the fitness function, and (b) evolving a "new" set of parameters, with a fitness function based on a non-orthogonal tight-binding method, which are better suited for Si cluster energetics as compared to the published S-W potential. Evolution is driven by a fitness function based on the energies and forces calculated for Si_n clusters ($n < 7$), and is able to predict accurate energies for minimum energy and deformed configurations of Si_n (n = 7, 8, 33) clusters, which were not used in the fitness function.

1 Introduction

Accurate molecular dynamics (MD) or atomistic simulation of reactive systems containing many atomic species is important for the conceptualization, design and testing of novel nanoscale materials, devices, systems and applications, and a broad range of physical and chemical phenomenon in other areas as well. Some well-studied processes, through MD simulations, include crack propagation in bulk materials [Yu, Kalia, Vashishta (1997)], thin-film deposition and etching [Srivastava, Garrison, Brenner (1991)], ion and cluster bombardment of solid-surfaces [Garrison (1992)], surface diffusion and reactions [Garrison, Kodai, Srivastava (1996)], and hetero-layer epitaxy, superlattices and quantum dots [Tsuruta, Omeltchenko, Kalia, Vashishta (1996)]. In the nanotechnology arena: physical and chemical characterization of carbon nanotubes and fullerenes, design and operation of molecular gears [Han, Globus, Jaffe, Deardorff (1997)], hinges, three-way junctions, and bearings have also utilized MD simulations using reactive dynamics of two and three species systems [Globus, et al. (1998); Srivastava, Menon, Cho (2001)]. However, as the system and device sizes continue to shrink and composition becomes more multi-species, there is a need for developing good quality reactive atomic force field functions that are not currently available.

For example, in the last twenty years more than 30 reactive atomic force field functions for Silicon (Si) have been developed by various groups. Only a few have survived the rigors of being used in MD simulation and comparison with the available experimental data [Stillinger, Weber (1985); Tersoff (1989); Bazant, Kaxiras (1997); Garrison, Srivastava (1995)]. The Stillinger-Weber (S-W) and Tersoff potentials were expanded to include multi-component systems such as Silicon-Germanium-Carbon by Tersoff, Si-H and Si-F extensions of S-W, and the very extensively used Tersoff-Brenner potential for hydro-carbon systems [Brenner (1990)]. The development of such functions has not extended towards other multi-component systems such as Carbon-Boron-Nitrogen for nanotechnology applications, Carbon-Halogen systems for etching processes, and biological systems containing nitrogen, phosphorus, sulfur, oxygen and hydrogen atoms. The realization has been that developing reactive multi-atomic force field functions is difficult and tedious and, thus, is rarely attempted.

There are two parts to developing atomic force field functions. First, finding an analytic functional form that reflects the physical and chemical nature of the atomic

[1] NASA Ames Research Center, CSC/NAS, Moffett Field, CA 94035

[2] Department of Physics and Center for Computational Sciences, University of Kentucky, Lexington, KY 40506-0045

species under consideration, and second, fitting parameters in a multi-dimensional space based on the data available from experiments or more accurate quantum mechanical calculations. Choice of a functional form is complex and has been investigated in detail [Bazant, Kaxiras (1997)]. Much of the tedium, however, lies in the process of parameterization and comparison with the observables available from other sources. In an ideal case, the cycle of choosing a functional form and parameterization of the force field function should be iterated until a reasonable convergence is achieved with the widest variety of experimental and computational data available for the atomic species under consideration. Doing this for multi-component systems would be extremely tedious because the parameter space that needs to be investigated is large and could be correlated in a complex way. For biological systems, a variety of atomic force field functions with parameters have been developed and are available in commercial packages. All of these are non-reactive in nature and not suitable for studying the nanotechnology-based materials, systems and applications described above.

While automatic development of a physics-based functional form for atomic systems is beyond the current capabilities of computer science, fitting parameters in a complex multi-dimensional space to a given data set is not. As typical computing resources continue to increase many-fold every year, we hypothesize that parameterization of complex inter-atomic potential functions can be automated by large genetic algorithm computations on cycle-scavenged desktop computers. This should allow atomic force-field developers to concentrate mainly on the functional form and gathering experimental and higher accuracy simulation data to drive the evolution and validate the results. The resulting process may enable the routine exploration of the individual functional forms for multi-atomic species systems as well as the iteration of the procedure until a good convergence with a wide set of available data is achieved.

Using genetic algorithms (GA) in the proposed scheme has two advantages. First, GA is geared towards sampling both the near-equilibrium (minimum energy) and far-from-equilibrium (energetically excited) configurations in the data-set, and second, thousands of independent JavaGenes GA trajectories can be run in an embarrassingly parallel manner in a non-homogeneous distributed computing resource based environment. The term JavaGenes refers to a general purpose GA code written in Java programming language that can be run on a variety of computing platforms [Globus, Lawton, Wipke (1999), Globus, Langhirt, Livny, et. al (2000)]. In this work, we demonstrate the above described concept by fitting parameters of the well established Stillinger-Weber (S-W) Si potential by (a) first reproducing the published S-W parameters with a fitness function based on energies and forces calculated using the published S-W potential, and by (b) evolving a "new" set of S-W parameters with a fitness function based on the energies and forces calculated by a non-orthogonal tight-binding method for a better description of Si cluster energetics and dynamics which were not possible earlier with the published S-W potential. As a test, the evolution was driven by the fitness function involving energies and forces calculated for Si_n ($n < 7$) clusters, and is able to predict accurate energies for minimum energy and deformed configurations for Si_n ($n = 7, 8, 33$) clusters, which were not used in the fitness function. In Section 2 details of the method are described, whereas in Sections 3 and 4 we discuss the GA fitting of the S-W functional form for Si clusters.

2 Method

In this section we describe implementation of the JavaGenes GA for massively parallel search of multi-parameter space for fitting reactive many-body atomic force field functions. The scheme exploits the CPU cycle scavenging technology useful for these kinds of simulations, and the emphasis is on a possible future automation of the entire procedure for the parameterization of complex functional forms for solid-state systems containing multi-atomic species. The basics and details of the JavaGenes GA, a massively parallel implementation using CPU cycles scavenged by the Condor [16] system are described first, and S-W force field function that is used as an example for testing and validation of the approach is discussed later.

2.1 Genetic Algorithm Approach for Fitting Molecular Potentials

There have been a few recent examples of using GA to find atomic interaction potential parameters for "non-reactive" approaches such as molecular mechanics (MM2) [Mohamadi, Richards, Guida, et. al. (1990)] for metal-organic complexes and Amber parameters for organic molecules. Parameters for force field functions for tripod metal compounds using GAs [Hunger, Beyreuther, Huttner (1996); Hunger, Beyreuther, Huttner, et. al. (1998); Hunger, Huttner (1999)] have been developed where the fitness function was based on crystal structure conformations. Cundari used a similar technique to develop force field parameters for Technetium (Tc) complexes [Cundari, Fu (2000)]. Instead of using differences in predicted atomic locations, they used the difference of energies predicted by quantum effective core potential calculations and MM2 calculated energies for evolving the parameters, and have found that GA evolved parameters could improve the predictive power of MM2 over parameters derived from quantum chemistry calculations by other techniques. Wang and Kollman have optimized Amber force field parameters for several organic molecules using GA and compared the results to a systematic search [Wang, Kollman (2001)]. The GA was found to be more efficient as long as at least three parameters were being optimized, and in some cases the GA also found better parameters. To date there has been no attempt to use GA for finding atomic force field parameters for reactive systems interacting with many-body force field functions where the parameter space is complex and multiply connected.

The GA seeks to mimic natural evolution's ability to produce highly functional objects. Natural evolution produces organisms, whereas the GA can produce sets of parameters, programs, molecular designs, and many other structures. Our GA, JavaGenes, employs the following algorithm (words in quotes are typical GA terminology):

1. Represent potential parameters with a set of floating point numbers; each set is called an ”individual”
2. Generate a ”population” of individuals with random parameters
3. Calculate the ”fitness” of each individual
4. Repeat
    - Randomly select ”parents” with a bias towards better fitness
    - Produce ”children” from the parents with either a
        - ”crossover” that combines parts of two parents into a child
        - or ”mutation” that modifies a single parent
    - Calculate the fitness of the child
    - Randomly replace individuals of less fitness in the population with the thus produced children
5. Until satisfied according to some minimal convergence criteria

The vast majority of CPU time is usually spent calculating the fitness function. The above is easy to implement but hard to use, and in general GAs are not guaranteed to find a unique or even a satisfactory solution, but often work well in practice. There are a wide variety of GA techniques, and the implementations use many “GA parameters” that can affect performance of the search procedure. Examples of GA parameters include population size and the mix of mutation vs. crossover operators. Thus, choosing a proper GA technique and parameters is a non-trivial problem. We solve this by randomizing the choice of GA parameters in appropriate ranges in many GA runs. The main features of JavaGenes used for fitting molecular force field functions is described next.

JavaGenes is a steady state tournament selection genetic algorithm. The tournament size is usually two. In tournament selection each parent is chosen by randomly selecting two individuals from the population and choosing the fittest to be the parent. After crossover or mutation produces a child, individuals to replace are chosen by an anti-tournament of size two. In an anti-tournament the least fit individual is chosen for replacement by a newly created child. Steady state means that there is only one population, parents are chosen from this population and children replace individuals in the same population. During GA-parameter randomization the tournament size is probabilistically two or one. A tournament size of one means that a random individual is chosen as the parent. Size one ’tournaments’ help avoid premature convergence.

Mapping the problem of finding parameters for molecular force field functions on to a GA scaffolding is done by representing the force field parameters as a ragged two-dimensional array of double precision floating point numbers. The first dimension represents the two- or three-body terms of the potential function, and the ragged second dimension represents the parameters. A ragged second dimension means that arrays of differing length are in the second dimension. For example, one second dimension array may hold the two-body parameters for Si, this would be of length five for S-W case, and another second dimension array of length three might hold the Si three-body parameters. Each parameter is assigned a set of limits within which it is allowed to evolve. The limiting values of the parameters are chosen from the physical interpretation of the contribution of the parameter to the force field function.

During evolution, JavaGenes uses two transmission operators to generate children from parents. These operators are: mutation and interval-crossover. Mutation requires only one parent, a copy of the parent is made and some of the potential parameters are randomly modified. The JavaGenes mutation operator takes two GA-parameters: the probability any single parameter will be mutated, and the width of the Gaussian distribution, around the mean parental value, from which the required change is chosen. Interval crossover requires two parents. The parental-values of a parameter determine the extremes of the range within which the child-value of the parameter is randomly selected. This range can be increased or decreased by a factor (based on a GA-parameter), and the child's value can be with-in or out-of-range (based on another GA-parameter).

The evolution of a GA population is guided by a fitness function. The GA fitness function must provide a fitness for any possible individual, no matter how bad, and distinguish between any two individuals, no matter how close they are. The fitness function for determining parameters for molecular force field functions compares energies and forces computed for a given set of atomic conformations using the evolving parameters with externally supplied energies and forces. The inherent advantage of GA over other techniques, therefore, is that one can use close to equilibrium (energetically minimized) as well as far from equilibrium (energetically excited) configurations. This is significantly different from the approaches built around fitting only the near equilibrium configurations. Specifically, JavaGenes uses three forms of fitness functions. The forms are: (i) root-mean-square (RMS) deviation from externally supplied energies, (ii) RMS of $|a\text{-}b|/(|a|+|b|)$ where a and b are the calculated and externally supplied energies, respectively, and (iii) RMS of $|c\text{-}d|/(|c|+|d|)$ where c and d are the calculated and externally supplied forces, respectively. The first is an accepted measure of deviation, but has problems when the absolute value of the supplied energy varies wildly. For example, energies at very small separation have very large values and can have excessive influence on determining the full force field function. In reality, much of the room temperature and reactive state behavior is determined by the energies near equilibrium or large separations. The form used in (ii) and (iii) always returns a value between 0 and 1, eliminating the scaling problem of the form used in (i). The form in (ii) and (iii), however, may exhibit poor behavior if the calculated and standard values are of opposite sign. All three forms are combined by applying each to different conformations and taking a weighted sum as the fitness function.

A fitness function can be no better than the externally supplied energies and forces. These could be obtained from either better accuracy *ab-initio* or tight-binding interactions or from experiments. In a single objective GA, as described in this work, we either use the values known from the S-W potential itself to demonstrate the efficacy of the GA technique, or the values obtained from a non-orthogonal quantum tight-binding description for a range of Si cluster configurations to predict the values of the clusters not included in the fitness procedure. In a multi-objective GA, which has been implemented but not used for the data in this paper, it will be possible to fit to the experimental values as well as values obtained from higher accuracy computational approaches simultaneously.

For example, Table 1 summarizes the terminology as well as the details of the GA-parameters used in each of the experimental runs. The GA-parameters throughout the description are chosen to determine the run conditions of GA jobs while the overall objective is to fit the molecular force-field-parameters of a given functional form. The original JavaGenes [Globus, Lawton, Wipke (1999)] GA code, written for finding pharmaceutical drug molecules, was modified to evolve force field parameters for a given functional. As an example, in this work, we focus on demonstrating the efficacy of the method in finding Si force field parameters for S-W functional form. In the future, we will use the developed technique to fit the more complex force field functions for multi-atomic species systems.

In the beginning, 30-100 single-workstation GA trajectories with identical GA-parameters (except the random number seed) for each force-field-parameter search were run with populations varying between 1000-3000. The GA-parameters that worked for one search (say, Si

Table 1 : GA Parameters

Conformations	Describes the atomic conformations used in the fitness function
Conformation sets	Several sets of conformations were used. The fitness value for each set was calculated and a linear combination of the fitness for each set was the final fitness. 39 “far wall” dimers evenly spaced from 0.5 - 1.728 angstroms 7 “near wall” dimers evenly spaced form 1.599 - 1.793 angstroms 44 “minimum” dimers evenly spaced from 1.793 - 3.183 angstroms 41 “tail” dimers evenly spaced from 2.407 - 3.7 angstroms All other clusters are randomized around the minimum energy as calculated by tight-binding ● 67 3-atom clusters ● 51 4-atom clusters ● 41 5-atom clusters ● 34 6-atom clusters
Target energies (and forces)	Describes the source of the energies and forces used in the fitness function. This was always either the Stillinger-Weber potential with published parameters or the Menon tight-bonding code.
Energy (and force) comparison	Describes the function used to compare energies and/or forces with the target energies and/or forces RMS of $\|a-b\|/(\|a\|+\|b\|)$ on wall and tail RMS elsewhere Each set of conformations generated a seperate value and these were summed to get the fitness. The near wall value was multiplied by 0.5 before summation.
Number of jobs	Number of separate, single-workstation jobs in the run = 1001
Popluation size	Size of the population = 100
Children per generation	Number generated for each generation = 2000
Number of generations	Number of generations = 200
Transmission operators	Mix of crossover and mutation transmission operators = [0-5] interval crossover, [1-30] mutation chosen at random
Interval crossover parameter	For the interval crossover transmission operator, the amount the interval between parental values of a parameter grew or shrank before chossing a random value within the interval = [0.3 to 3]
Mutation frequency	The probability that any one GA-parameter was mutated by the mutation transmission operator = [0.1 to 0.9]

Mutation standard deviation	The mutation operator modified GA-parameters by choosing randomly from a Gaussian distribution centered on the parental value. This value is expressed as a fraction of the allowed interval for a GA-parameter = [0.1 to 0.9].
Stillinger-Weber parameter varies from	The range within which a Stillinger-Weber parameter could vary. Note that the interval crossover operator could be set to ignore this interval A,B,alpha, gamma = [-100 to -50] to [75 to 150] p,q = 0 to [12 to 24] C, gamma = 0 to [3 to 8]
Immigrants	When searching for both the two- and three-body parameters, JavaGenes sometimes initialized the population with two-body values taken from a run that focused on the two-body parameters. 25% of jobs initial population started with best two-body evolved parameters. 25% of jobs half of initial population started with best two-body evolved parameters.

GA-parameter values placed between brackets, ”[” and ”]”, indiicate that the value was chosen randomly within limits. For example, [0.1-0.9] inicates that a GA-parameter was randomly chosen for each job between 0.1 and 0.9 inclusive.

dimers in the fitness function) would fail in a similar search for a different system. The alternate technique of using approximately a thousand trajectories with randomized GA-parameters and smaller populations (100) worked very well for all the systems attempted. As stated above, we first reproduced the Stillinger-Weber (S-W) results using S-W small cluster energies derived from the published parameters in the fitness function. This shows that the method can find the global, not just the local, minimum. Then, using the same small Si$_n$ clusters (n < 7), we found a “new” set of parameters using the energies and forces supplied by a quantum non-orthogonal tight-binding method of Menon and Subbaswami (1993) that showed good results for small Si$_n$ clusters (n > 6) and Si$_{33}$ clusters that had both tetrahedral and under- over-coordinated Si atoms in the system.

2.2 An Example Molecular Force-Field Function: Stillinger-Weber (S-W) Potential

The above developed approach is for fitting parameters of molecular force field functions for complex multi-atomic species systems not available so far. Choosing a functional form to describe such a complex system with reasonable accuracy is an involved process and will be attempted in the future. The focus of this work is on establishing the GA technique for finding the force field parameters for a given functional form. We have chosen the S-W functional form as an example and fitted the parameters using the GA approach in the two cases as described above. In this section we briefly discuss the S-W functional form and the parameters that need to be evaluated using the GA approach.

The S-W molecular potential expresses the total energy of a given configuration in terms of the sum of two- and three-body contributions to the energy as a function of the atomic positions in the configuration:

$$E = \sum_{\substack{i,j\\ i<j}} v_2(i,j) + \sum_{\substack{i,j,k\\ i<j<k}} v_3(i,j,k) \qquad (1)$$

where E is total interaction energy, i,j,k indicate individual atoms, and v is the interaction energy of n atoms.

To reduce computation, reactive potentials often have a cutoff function which forces each term to zero at large atomic separations. This converts the problem from

O(n^3) to O(n) since only near neighbors need be considered. The S-W potential used in this paper only considers two- and three-body terms and has an exponential cutoff on both the terms. The terms are:

$$v_2(i,j) = A(Br^{-p} - r^{-q})c_1 \tag{2}$$
$$c_1 = e^{\frac{C}{r-a}}; \qquad r < a$$
$$c_1 = 0; \qquad r \geq a$$

where r is the i,j inter-atomic distance and all other values are adjustable parameters.

$$v_3(i,j,k) = \alpha + \lambda(\cos\theta - \cos\theta_0)^2 c_2 \tag{3}$$
$$c_2 = e^{\frac{r}{r_{i,j}-a_1} + \frac{r}{r_{j,i}-a_1}}; \qquad r_{i,j} < a_1 \cap r_{j,k} < a_2$$
$$c_2 = 0; \qquad r_{i,j} \geq a_1 \cup r_{j,k} \geq a_2$$

where r$_{ij}$and r$_{jk}$ are the two inter-atomic distances, theta is the angle and all other values are adjustable parameters. The parameters a, a$_1$, and a$_2$ defining the cut-off distance on the two- and three-body terms are not evolved because their choice is determined by the physical and chemical considerations in the system. The typical cutoff distances are chosen within first and second neighbor distances so as to keep the numerical efficacy of the short range reactive potentials. Lastly, the preferred bond angle θ_0 is also not evolved since it is readily available from experiment and theoretical considerations (θ_0 is the tetrahedral angle in solid-state Si).

The functional form should reflect the physics of the system of interest, and parameters should allow the form to fit the available data, although sometimes specific parameters have specific physical meaning. The relevant data include energies and forces for various atomic conformations calculated by higher accuracy methods, bond lengths, angles for energy minimized structures, bulk lattice constants, elastic and vibrational properties, and a host of other experimental data. Most of the tedium in multi-species reactive potential function development is in parameter fitting. Thus, if the multi-dimensional fit to the complex parameter space can be automated, then rapid development of broadly applicable potentials may be enabled by iterating the coarser grain procedure on the choice of functional forms as well.

2.3 CPU Cycle Scavenging System: Condor

For the current work, we used the Condor [Litzkow, Livny, Mutka (1988)] cycle scavenger running on about 350 SGI and Sun machines at the NASA Advanced Supercomputing (NAS) Division [www.nas.nasa.gov]. Each workstation runs a daemon that watches user I/O and CPU load. When a workstation has been idle for 2 hours, a job from the batch queue is assigned to the workstation and will run until the daemon detects a keystroke, mouse motion, or high non-Condor CPU usage. At that point, the job is removed from the workstation and placed back on the batch queue. The job eventually runs again, although probably on a different machine. Typically, between 100-250 NAS machines are available for batch processing through the NAS Condor pool at any one time. Although the NAS Condor pool supplies substantial processing power, it is by no means the largest cycle-scavenging compute facility. The best-known cycle-scavenging computation is seti@home [setiathome.ssl.berkeley.edu], that typically uses more than 3 million computers to provide about 23 teraflops/sec.

While cycle-scavenging systems can supply huge amounts of CPU, they are restricted to embarrassingly parallel problems with minimal I/O requirements. Many important problems fit within these restrictions, including parameter studies, Monte Carlo simulations, and GAs. For example, part of the data for this paper was generated by running 1000 ~8 hour genetic algorithm (GA) jobs with randomized GA-parameters. This procedure can use hundreds of processors with no inter-process communication and minimal disk I/O. As a result, typically 2000+ CPU hours of computation is routinely accomplished overnight without purchasing any new hardware. The results described below are reproducible only in a statistical sense - although repeated tries of the same runs give similar results. The runs are not exactly repeatable because of permitted variations in IEEE floating point arithmetic combined with cycle-scavenging in a heterogeneous environment. The variation, however, appears to be well within the range of error associated with the accuracy of atomic force field functions.

While only some computations can use cycle-scavenging, those that can, such as JavaGenes, need not be very concerned with efficient use of CPU cycles since the vast majority of CPU-cycles on the vast majority of all desktop computers do absolutely nothing other than wait for user input. Thus, even the most inefficient cycle-scavenging computation makes better use of the available resources so long as each desktop computer rapidly responds to user input (mouse motion or keystroke).

3 Results

As an example of fitting parameters of known and well established molecular force function with the above described methodology we chose the Stillinger-Weber (S-W) potential described above. First, as validation, we use the published S-W potential calculated energies and forces of small Si_n ($n < 7$) clusters in the fitness function, and compare the results in the case of Si_n ($n > 6$) clusters that were not used in the fitness function. Second, as a test of the approach, we find "new" GA evolved S-W potential parameters where only the functional form was assumed to be known, and the fitness function was described by energies and forces of small Si clusters computed from the non-orthogonal tight-binding scheme of Menon and Subbaswamy. The fitness function based on small Si_n clusters gives GA evolved potential parameters that describe the energetics of both small and large Si_n clusters rather well.

3.1 Validation: evolution and comparison with published S-W values

In the first attempt, the two-and three-body terms were fitted separately and sequentially, with an attempt to find out if the feeding of the first fitted two-body term (with a Si dimer) in the GA job for a three-body term (for a Si trimer) facilitates or accelerates the fitting procedure. No such facilitation was observed. Never the less, some important observations were made. For example, in Table 2a, we show the GA evolved parameters for the two-body term based on a fitness function spanning the energies of 100 Si dimers equally spaced within the range of 0.5 to 3.7 Ang. At first glance, the evolved parameters seem to be incorrect. However, it turns out that C is nearly correct and p and q are (approximately) reversed. This is because p and q are related through dependence on A and the GA evolution has essentially performed an algebraic operation during the fitting procedure. The equivalence between the published and evolved expressions is shown in Table 2b. The comparison of energy and forces obtained from the evolved parameters and the published parameters in Figure 2 shows a good fit in the entire range.

The parameters for two- and three-body terms together using 1000 GA jobs with a fitness function based on the energies of 2-6 atom Si clusters were evolved. The two-atom clusters or dimers were the same ones used in the two-body GA jobs described above. The minimum energy configurations for 3 to 6 atom Si clusters were first generated using the generalized tight-binding molecular dynamics (GTBMD) method of Menon and Subbaswamy. Using the minimum energy configurations as seed, the rest of the hundreds of conformations were generated by random displacements of atomic coordinates around minimum energy configurations. The fitness function was based on the energies and forces computed for the members of the population using the published S-W potential. In some jobs, part of the initial population was set to the best two-body parameters of the independently evolved two-body parameters. The two-body parameters were, however, allowed to evolve further with the rest of the three-body parameters, and no noticeable difference was observed in using this strategy as compared to the case where the full search did not assume any previous knowledge of the best two-body parameters. Table 3 shows the most fit GA evolved parameters as compared with their published value in the original S-W potential.

Figure 1 compares the energies of Si clusters as calculated by S-W potential with GA evolved parameters with those computed by using the published parameters in two cases. First, in Figure 1(a), we show the comparison for Si_n clusters with $n \leq 6$, i.e., the clusters used in the fitting procedure. Second, in Figure 1(b), the comparison is shown for Si_n clusters for n = 7, 8, i.e., clusters not used in the fitting procedure. The figure shows the comparison of the energies in the full range of the configurations, i.e., the minimum energy configurations as well as the configurations generated by the random displacements of the atomic positions around the minimum energy positions. The comparison shows a good fit in both cases. The deviation is found to be within a few tens of kcal/mol**.**

Figure 2 compares the energy and force curves generated by the two-body term of the published and GA evolved S-W potential. The fit is very good but not quite exact. The GA techniques are often efficient in getting close to the desired values but are not for the final refinement towards the exact fit. For materials physics and chemistry, where the data for the fitness function is a general but not exact reflection of the reality, this limitation is not serious and perhaps even an advantage. The figures 3 (a) and (b) show contour plots of the comparison of the en-

Table 2 a: Si two-body parameters

Parameters	S-W Published value	Evolved with S-W Fitness Function
A	7.0495	-4.21
B	0.602	1.67
C	1.0	1.01
p	4.0	-0.05
q	0.0	4.01

Table 2 b: Parameters of Table 2a rewritten for comparison

Energy $(r) = A(Br^{-p} - r^{-q})$	Published	Evolved
Inital form	$7(0.6r^{-4} - r^{0})$	$-4.2(1.67r^{-0.05} - r^{-4.01}$
with A distributd	$\mathbf{4.2r}^{-4} - 7r^{0}$	$-7.01r^{-0.05} + \mathbf{4.2r}^{-4.01}$

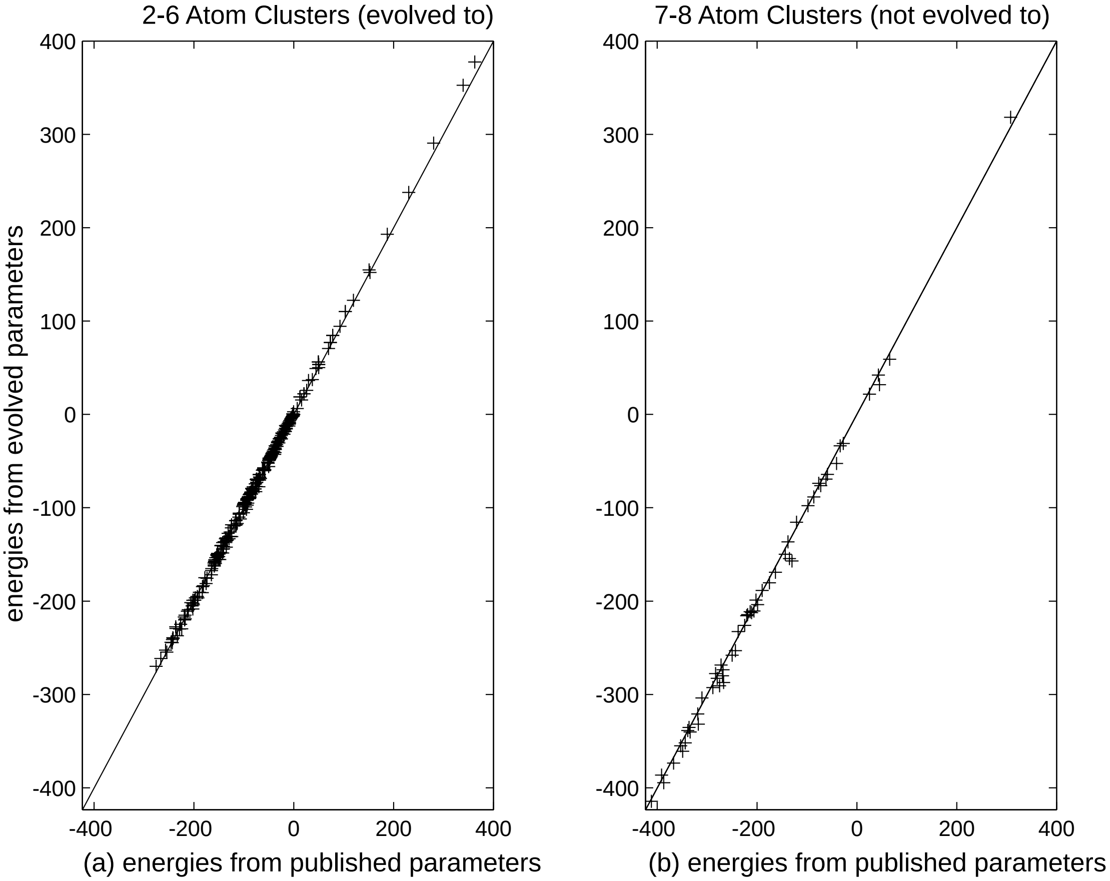

Figure 1 : Comparison of energies calculated for Si2-8 clusters using the published and evolved S-W parameters. Each cross represents a cluster. The horizontal/vertical axes are the energies calculated using the published/evolved parameters in kcal/mol. Crosses on the diagonal line are a perfect fit: (a) shows data for clusters (2-6) that were used in the fitness function, and (b) shows results for clusters (7,8) that were not used in the fitness function.

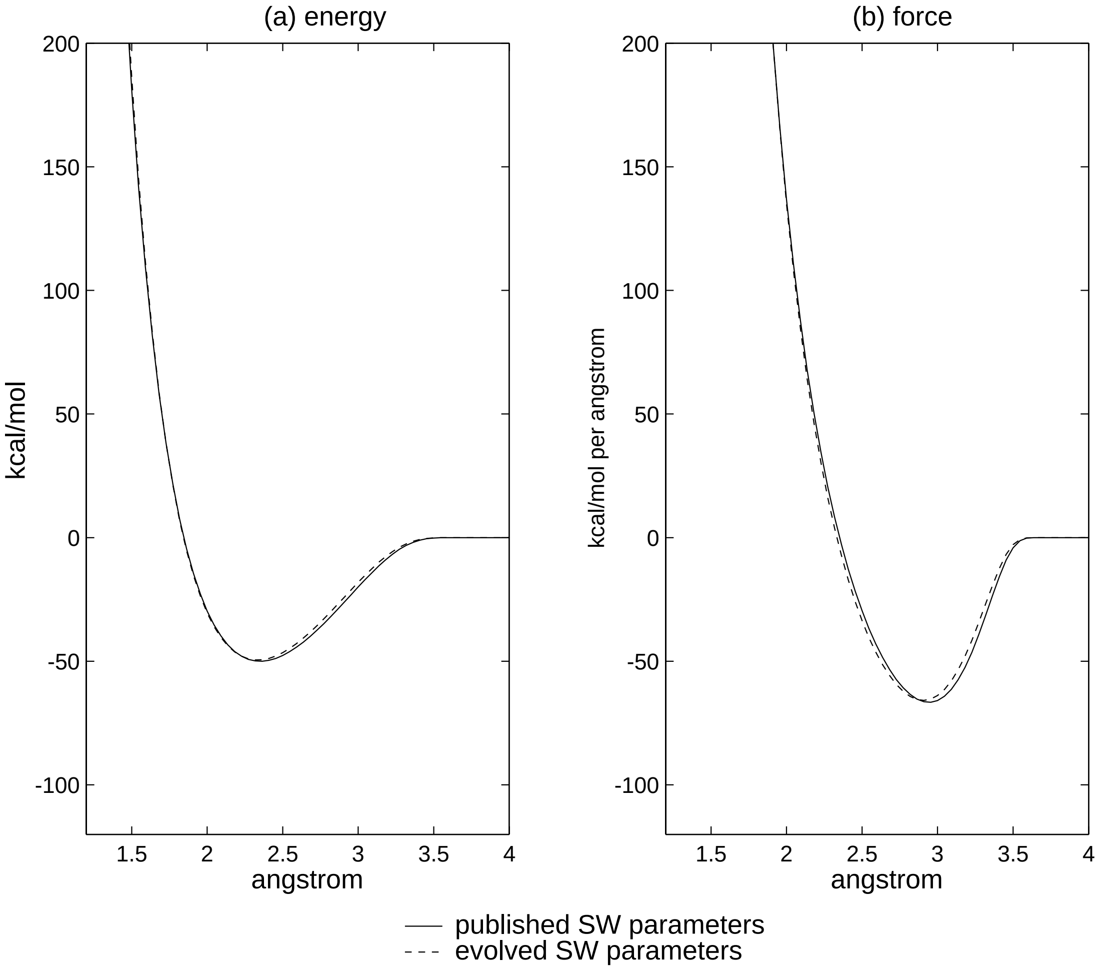

Figure 2 : Comparison of 2-body energies and forces calculated using published and evolved S-W parameters. Parameters were evolved using a fitness function with Si2 cluster energies calculated by S-W with published parameters. The solid lines represent values calculated using the published parameters. The dashed lines represent values calculated using evolved parameters.

Table 3 : Si parameters for S-W fit

Parameters	S-W Published Value	Evolved with S-W Fitness Function
A	7.0495	-4.51
B	0.602	1.68
C	1	1.06
p	4	0.015
q	0	4.066
Alpha	0	-1.68
Lambda	21	30.5
Gamma	1.2	1.289

ergies for the three-body term where both the angle and the bond lengths of a 3 atom Si cluster are varied within a possible range. The comparison shows a good fit but not as close as that for the two-body term in Figure 2. This is because the three-body term tends to contribute much smaller energies towards the total energy of a cluster, as compared to the contribution of the two-body term, which dominates the value of the fitness function during the evolution. It is conceivable that in a multi-objective GA evolution, in the future, we may use a physical observable in the data that is more sensitive to the contribution of the three-body term as compared to that of the two-body term.

3.2 *Test: evolution of new S-W Potential parameters for small and large Si clusters*

Having validated JavaGenes by reproducing the two- and three-body parameters published by Stillinger-Weber, and comparing the energetics of Si clusters not used in the fitting procedure, in this section, we describe evolving “new” SW parameters suitable for describing the dynamics of small and large Si clusters. This is done by constructing the fitness function by energies and forces calculated by a non-orthogonal tight-binding quantum description of Si interactions by Menon and Subbaswamy (1993), which has been previously shown to describe the energetics and dynamics of small and large Si clusters rather well.

The generalized tight-binding scheme differs from the conventional orthogonal tight-binding schemes in that explicit use is made of the nonorthogonality of the orbitals. This allows for proper accounting of local environments. The method has been successfully used for silicon [Menon, Subbaswamy (1997)], germanium [Menon, Condens (1998)] and carbon [Menon, Richter, Subbaswamy (1996)] systems to give good agreement in the range all the way from a few atoms to the condensed solid. Additionally, the vibrational frequencies for the dimer and also for the bulk structures at various symmetry points are in excellent agreement with experiment.

Parameters for the Si dimer were evolved using non-orthogonal tight-binding energies and forces in the fitness function. The S-W functional form has been constructed to give energies and forces going to zero at the cut-off distance and beyond because it is a short range potential. No such restriction, however, is imposed on the energies and forces computed from the quantum tight-binding method. At or near the minimum energy configuration and at short separations, therefore, we use energies in the fitness function. For long separations far from equilibrium the fitness function was based on using the forces computed by the quantum tight-binding method. A total of 1001 GA trajectories were run with randomized GA-parameters and the best fit values of the fitted S-W potential parameters are listed in Table 4. The parameters evolved with the fitness function based on tight-binding energies and forces appear quite different from those evolved with the fitness function based on S-W energies and forces. However, given a functional form for the molecular force field, there is no unique set of parameters that will be suitable for describing the resultant energetics and dynamics of the system. The real test, however, is how the energies and forces computed by the “new” set of parameters compare with those computed by the published set of parameters and how these compare with the energies and forces for a broader set of applications. The comparison of the energies and forces in the two cases of GA evolution is shown in Figure 4(a,b).

The “new” evolved parameters match the tight-binding data much better than the published parameters for all configurations. The major difference between the S-W evolved parameter energies and the tight-binding energies is at separations far from equilibrium (> 3 angstrom) where the S-W form is required to go to zero at the cut-off distance and beyond. In this region we have used forces, instead of energy in the fitness function. The force curve shows a smooth matching between the S-W evolved curve and the tight-binding evolved curve between 3 angstrom and the cut-off distance. Having found a good match for the two-body terms, the parameters for the full potential were evolved with the fitness

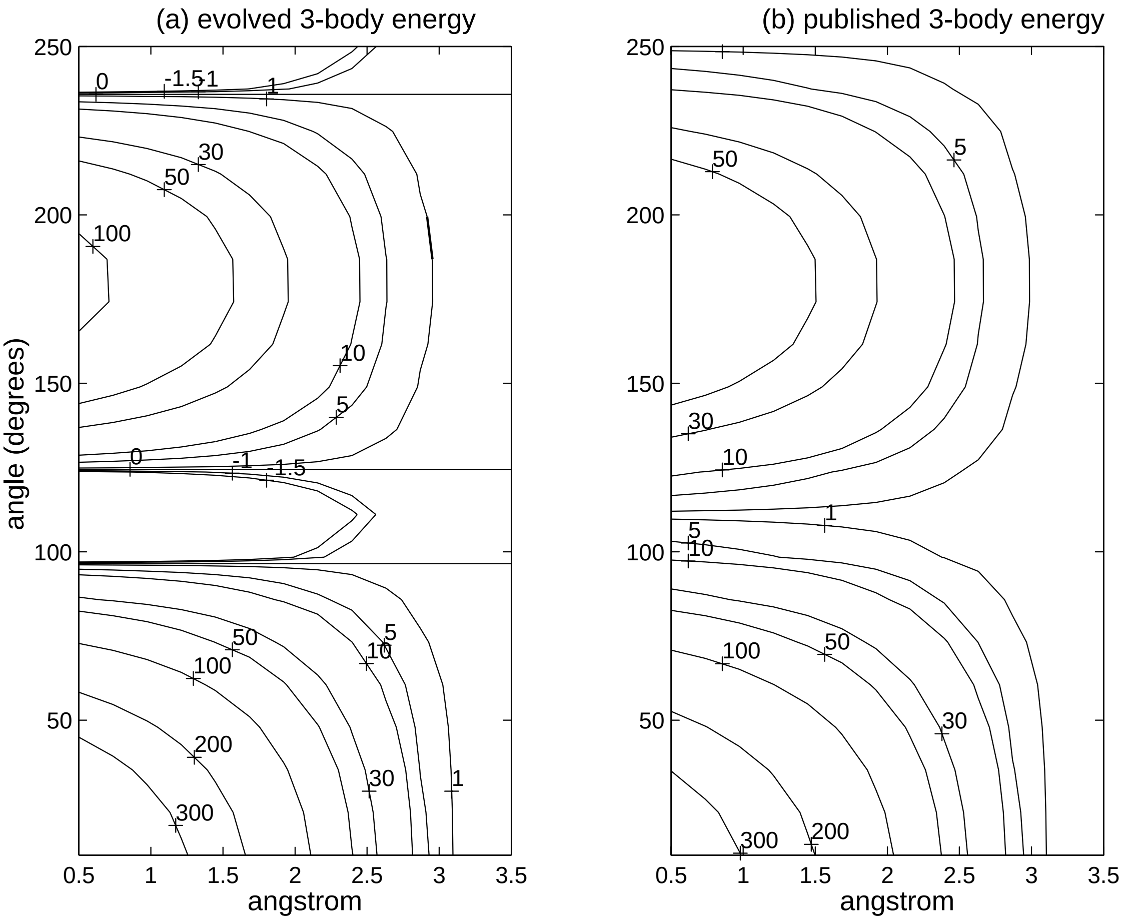

Figure 3 : Comparison of the 3-body energies calculated using the published and evolved S-W parameters. For each 3-body calculation, both bond lengths are kept equal: (a) shows the energies calculated using the evolved parameters, and (b) shows the energies calculated using the published parameters. Note that the minimum around 109.5 degress in (a) is below –1.5 kcal/mol.

Table 4 : Parameters for S-W published, S-W evolved, Tight-binding evolved cases

Parameters	S-W Published Values	Evolved with S-W fitness functions	Evolved with tight-binding fitness function
A	7.0495	-4.21	-0.66
B	0.602	1.67	14.23
C	1	1.01	1.48
p	4	-0.05	-2.50
q	0	4.01	18.67
Alpha	0	-1.68	11.7
Gamma	21	30.5	10.9
Lambda	1.2	1.289	1.38

function based on the tight-binding derived energies and forces of 2-6 atom Si clusters. The Si dimers were handled as above, and the seed for the 3-6 atom Si clusters were the minimum energy configurations computed by quantum tight-binding method. The remaining configurations were generated by random displacements of the atomic positions near minimum energy configurations. A total of 1001 GA jobs were run with randomized GA-parameters, and the role of one of these parameters was to import the previously evolved two-body potential parameters as a starting point for part or all of the population.

The best of these sets not only matched the energies of the 2-6 atom Si clusters, but also of the 7 and 8 atom Si clusters near minimum energy, and the configura-

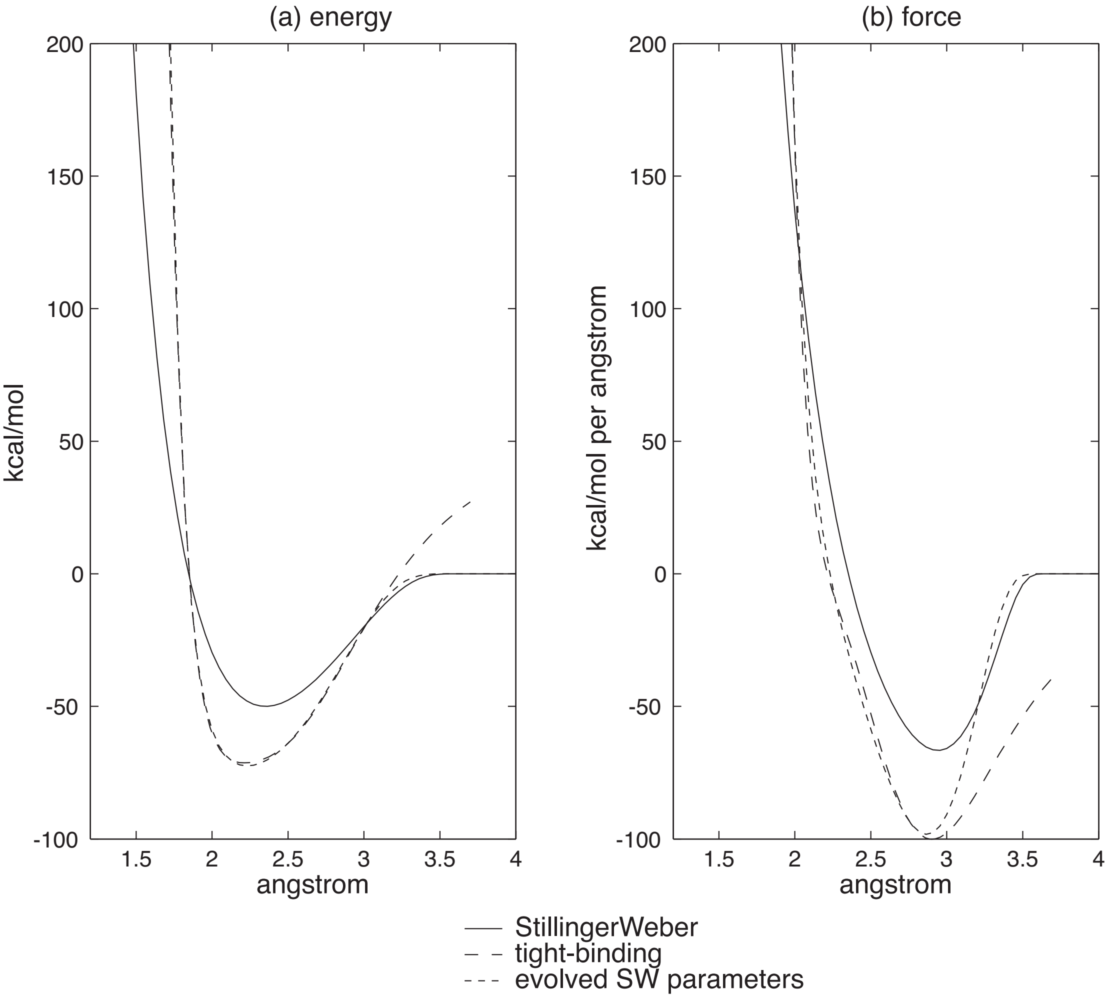

Figure 4 : Comparison of 2-body energies and forces calculated using tight-binding; and published and evolved S-W parameters. The long dashed lines represent values calculated by the tight-binding code. The solid lines represent values calculated using the published parameters. The short dashed lines represent values calculated using evolved parameters.

tions generated by random displacements of atomic coordinates around their minimum energy configurations. These are shown in Figure 5 (a) and (c). The match is very good for 2-6 atom Si clusters, as might be expected, because the energies and forces for these clusters were used in the fitness function. The comparison of the energies of 7 and 8 atom Si clusters, which were not used in the fitting procedure, also show good results suggesting that the approach is transferable. Figure 5 (b) and (d) shows the comparison of tight-binding energies with energies calculated from the original S-W parameters. The fit is much worse than for the evolved parameters.

A comparison of two-body energies and forces in the three cases, with the energies and forces computed by the published S-W parameters and tight-binding, is shown in Figure 6. The energies and forces generated by the S-W evolved parameters are somewhat different from the energies and forces computed by the evolved parameters with fitness function based on dimers. Specifically, the minimum energy is lower. This is explained by examining the energies of the three-body S-W term in isolation (Figure 7 which also shows the published SW parameter results for comparison). We found that the shallow well depth near minimum in the two-body term is compensated by the larger contribution from the three-body term. These are the energies from the S-W three-body

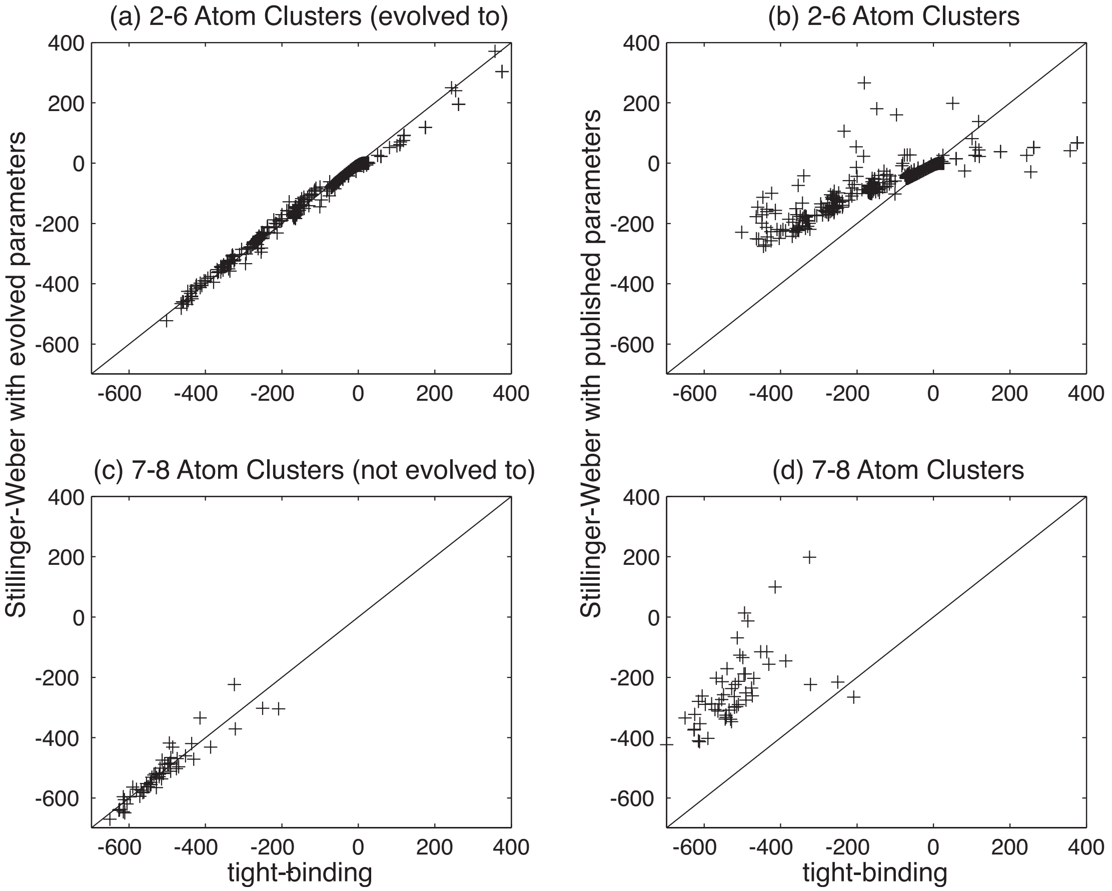

Figure 5 : Comparison of energies calculated for Si2-8 clusters using tight-binding; and published and evolved S-W parameters. Parameters were evolved using a fitness function with Si2-6 cluster energies calculated by the tight-binding method. For (a) and (c) the vertical axis is the energy calculated using evolved parameters in kcal/mol. For (b) and (d) the vertical axis is the energy calculated using the published parameters. Horizontal axes are the energies calculated by the tight-binding method. Figures (a) and (b) show data for clusters (2-6) that were used in the fitness function. Figures (c) and (d) show results for clusters (7-8) that were not used in the fitness function.

term alone, and no such comparable separation of terms exists in the tight-binding approach. This is natural because the cutoff for the S-W three-body term depends on the lengths of the two involved bonds, and the GA evolution has moved some of the functionality of the two-body term into the three-body term and sacrificed on the energy of an isolated Si dimer configuration. We note that at long separations, the evolved parameters and the force field functions show little preference for any angle in the configuration because the energetics is dominated by the two-body terms. This is natural as an isolated Si atom start to approach a Si dimer, at large distances the approach pathway is equally favorable from all directions. The angular part starts to dominate the configuration as the atoms are pulled in closer at short distances.

Finally, a more rigorous test of the fitted tight-binding evolved parameters is to check the energies of even larger Si clusters, with under and over coordinated Si atoms, such as Si_{33} with the newly fitted silicon S-W potential parameters. The silicon coordination and bonding in larger Si clusters such as Si_{33} are different from the coordination and bonding in small Si_n ($n < 7$) clusters because Si atoms are present in the surface configurations as well as in bulk configurations together with under- and over-coordinated Si atoms. The energy minimized structures of Si_{33}, calculated by the tight-binding molecular dynamics method, have been reported in the literature and are found to be stable under dynamic condi-

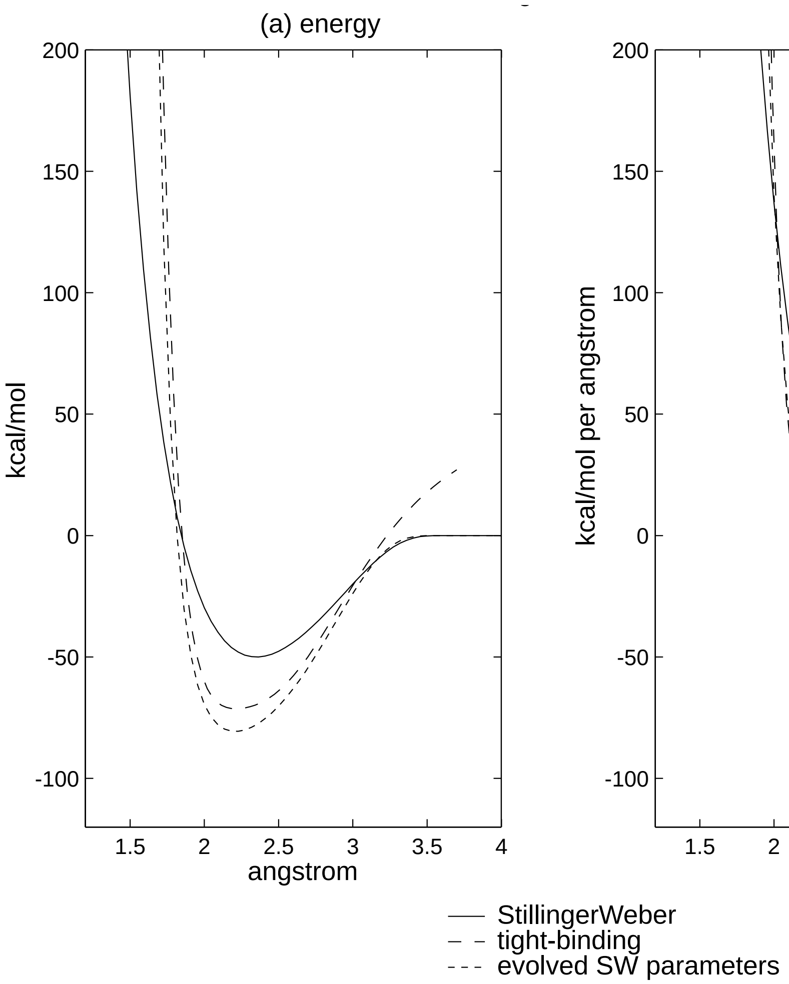

Figure 6 : Comparison of 2-body energies and forces calculated using tight-binding; and published and evolved S-W parameters. Parameters were evolved using a fitness function with Si2-6 cluster energies calculated by the tight-binding method. The long dashed lines represent values calculated by the tight-binding code. The solid lines represent values calculated using the published parameters. The short dashed lines represent values calculated using evolved parameters.

tions. Figure 8(a) compares the evolved S-W energies with those from the tight-binding energies of the near equilibrium and/or deformed Si_{33} clusters randomized around the equilibrium configurations. As with other results, the fit is closest at lower energies (within 200 kcal/mol) and is within 500 kcal/mol for heavily deformed configurations, although most configurations are much closer (mean 88, std 98, max 469). However, Figure 8(b) shows that S-W with the original parameters fits the tight-binding energies much less well (mean 500, std 327, max 1073).

4 Comments

Given a functional form for molecular force field functions, we have shown that genetic algorithms (GA) show promise for automating the task of fitting parameters over a complex range of configurations using large amounts of otherwise unused compute cycles in a distributed non-homogeneous computing environment. The GA fitness function is based on energies and forces of atoms and clusters near as well as far from equilibrium configurations. Therefore, it is possible to include a rather complete sampling of the configuration space as compared to the methods that are based mainly on the energies of the near equilibrium configurations. Specific choosing

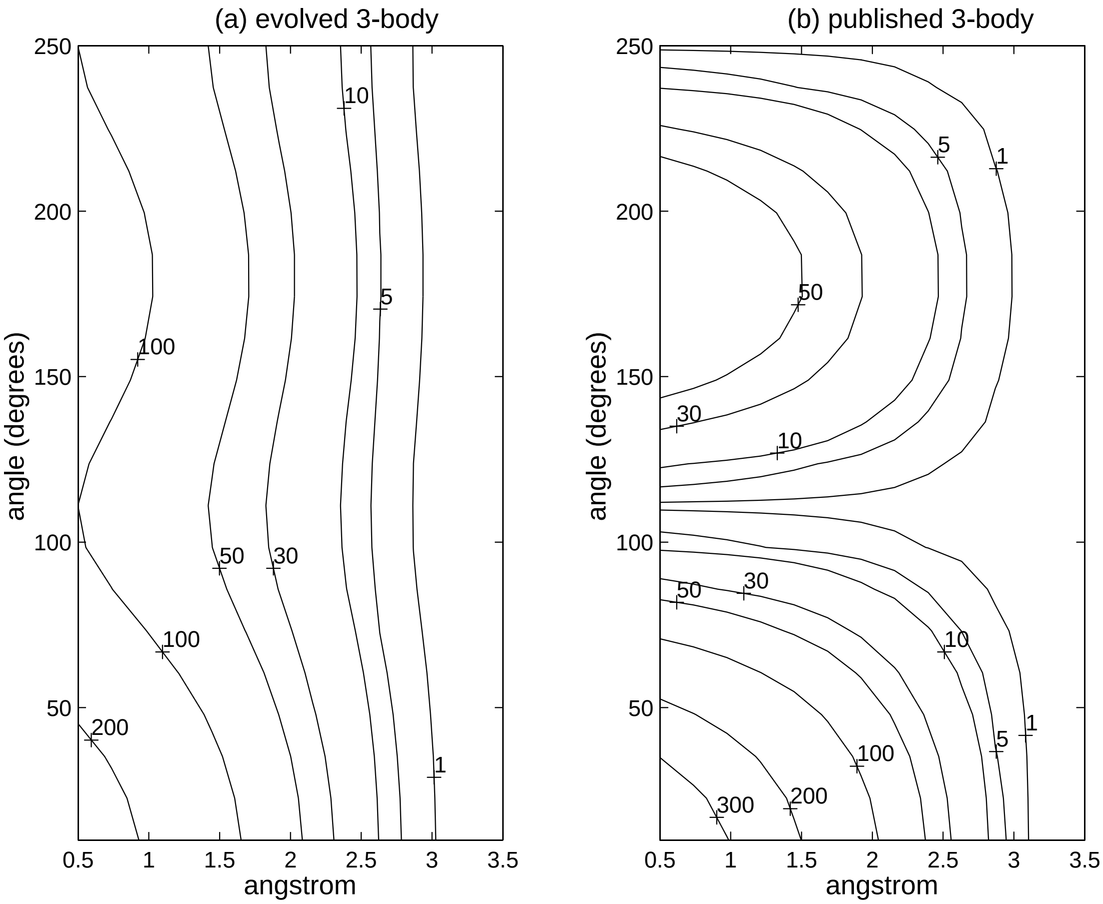

Figure 7 : Comparison of the 3-body energies calculated using the published and evolved S-W parameters in the similar scenario as in Fig. 6. For each 3-body calculation, both bond lengths are kept equal: Figure (a) shows the energies with the evolved parameters, and (b) shows energies with the published parameters.

of GA-parameters during the fitting procedure was found to be very time consuming and involved. The chosen GA parameters would work in some cases and would not work in other cases. However, taking advantage of CPU cycle scavenging through Condor, we found that the randomization of the GA-parameters within suitable ranges over many runs is an effective strategy.

As an example, in this work, we have demonstrated and validated the approach by finding parameters of the S-W Si potential in a variety of scenarios. Using the energies and forces computed by the published S-W Si potential in the fitness function, first, we have shown that JavaGenes is capable of reproducing the published S-W parameters and the energy and force curves – to very high precision. Using the energies and forces computed by a non-orthogonal tight-binding quantum description, for small Si$_n$ (n $<$ 7) clusters, in the fitness function we have found a "new" set of Si parameters for the S-W functional form. The "new" set of GA evolved Si S-W parameters, not only reproduce the energies of the small Si$_n$ (n $<$ 7) clusters used in the fitting procedure but also of Si$_n$ (n = 7,8,33) which were not used in the fitting procedure. The bonding and coordination in the largest cluster is significantly different from that of the small Si clusters (that were used in the fitting procedure) and yet the energies are reproduced well in comparison with the non-orthogonal tight-binding energies. We believe that this is primarily because both the near and far from equilibrium configurations for small Si clusters are used in the fitness function. The distorted Si clusters, even for small Si clusters, are able to include the effect of under- and over-coordination in the fitting procedure and the resulting "new" parameters give good results for Si$_n$ (n = 7,8,33). Work on extending the approach to

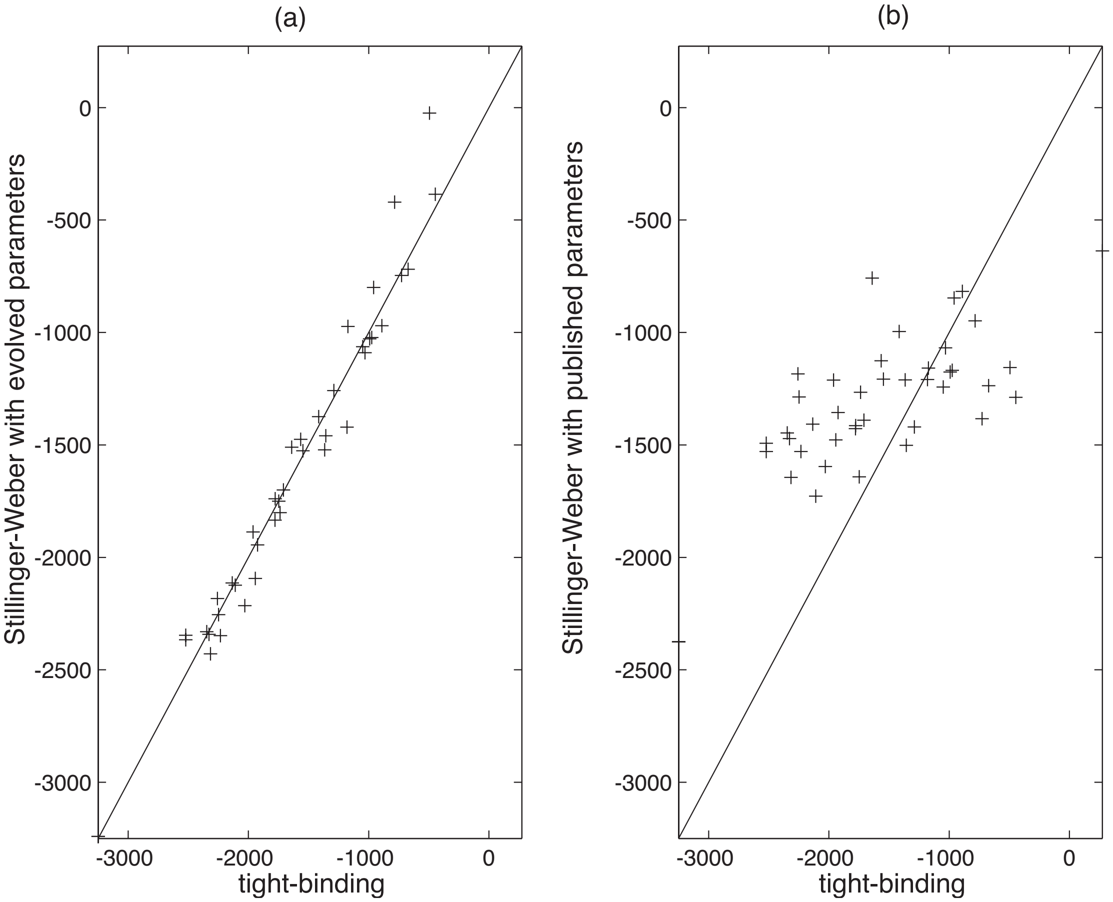

Figure 8 : Comparison of energies calculated for Si_{33} clusters using tight-binding; and published and evolved S-W parameters in the similar scenario as in Fig. 8.

evolve parameters for multi-component atomic systems, and generalizing the implementation to include experimental data (via a multi-objective GA) in the procedure, is in progress and will be published elsewhere.

Acknowledgement: We thank NASA's NAS supercomputing facility for funding and computational support, Eric Langhirt for maintaining the NAS Condor Pool and solving systems issues as they arose, and Charles Bauschlicher and Sandra Johan for assistance. We thank NASA's NAS supercomputing facility for funding and computational support through the CICT program ITSR project. Part of this work (AG and DS) is supported by NASA contract 704-40-32 to CSC.

References

Bazant, M. Z.; Kaxiras, E. (1997): Environment - dependent interatomic potential for bulk silicon. *Physical Review B*, vol. 56, no. 14, pp. 8542 - 8551

Brenner, D. W. (1990): Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. *Physical Review B - Condensed Matter*, vol. 42, no. 15, pp. 9458 - 9471

Cundari, T. R.; Fu, W. T. (2000): Genetic algorithm optimization of a molecular mechanics force field for technetium. *Inorganica Chimica Acta* 300-302, pp.113-124

Garrison, B. J. (1992): Molecular dynamics simulations of surface chemical reactions. *Chemical Society Review*, vol. 21, pp. 155 - 162

Garrison, B. J.; Srivastava, D. (1995): Potential energy surfaces for chemical reactions at solid surfaces. *Annual*

Review of Physics and Chemistry, vol. 46, pp. 373-394

Garrison, B. J.; Kodali, P. B.; Srivastava, D. (1996): Modeling of surface processes as exemplified by hydrocarbon reactions. *Chemical Reviews* 96, pp. 1327 - 1241

Globus, A.; Bauschlicher, C.; Han, J.; Jaffe, R.; Levit C.; Srivastava , D. (1998): Machine phase fullerene nanotechnology. *Nanotechnology*, vol. 10, no. 2, pp. 192 - 199

Globus, A.; Lawton, J.; Wipke, T. (1999): Automatic molecular design using evolutionary techniques. *Nanotechnology*, vol. 10, no. 3, pp. 290-299

Han, J. ; Globus, A.; Jaffe, R.; Deardorff, G. (1997): Molecular dynamics simulation of carbon nanotube based gears. *Nanotechnology*, vol. 8, no. 3, pp. 95-102

Hunger, J.; Beyreuther, S.; Huttner, G.; Allinger, K.; Radelof , U.; Zsolnai, L. (1998): How to derive force field parameters by genetic algorithms: modeling tripod-Mo(CO)3 compounds as an example . *European Journal of Inorganic Chemistry*, pp, 693-702

Hunger, J.; Huttner, G. (1999): Optimization and analysis of force field parameters by a combination of genetic algorithms and neural networks. *Journal of Computational Chemistry*, vol. 20, pp. 455-471

Litzkow, M.; Livny, M.; Mutka, M. W. (1988): Condor - a hunter of idle workstations. *Proceedings of the 8th International Conference of Distributed Computing Systems*, pp. 104-111

Menon, M.; Subbaswamy, K. R. (1993): Non-orthogonal tight-binding molecular-dynamics study of silicon clusters. *Physical Review B*, vol. **47**, no. 19, pp. 754-759

Menon, M. ; Richter, E.; Subbaswamy, K. R. (1996): Structural and vibrational properties of fullerenes and nanotubes in a non - orthogonal tight - binding scheme. *Journal of Chemical Physics*, vol. 104, pp. 5875 - 5882

Menon, M.; Subbaswamy, K.R. (1997): Non - orthogonal Tight - Binding Scheme for Silicon with Improved Transferability. *Physical Review B*, vol. 55, pp. 9231 - 9234

Srivastava, D.; Garrison, B. J.; Brenner, D. W. (1991): Modeling the growth of semiconductor epitaxial films via nanosecond time - scale molecular dynamics simulations. *Langmuir*, vol. 7, pp. 683 - 692

Srivastava, D.; Menon, M.; Cho, K. (2001): Computational nanotechnology with carbon nanotubes and fullerenes. *CMES: Computing in Science and Engineering*, July - August 2001, pp. 42 - 55

Stillinger, F. H.; Weber, T. A. (1985): Computer simulation of local order in condensed phases of silicon. *Physical Review B*, vol. 31, no. 8, pp. 5262-5271

Tersoff, J. (1989): Modeling solid - state chemistry: interatomic potentials for multicomponent systems. *Physical Review B*, vol. 93, no. 8, pp. 5566 - 5568

Wang, J.; Kollman, P. A. (2001): Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry*, vol. 22, no. 12, pp. 1219 - 1228

Yu, J.; Kalia, R.;Vashishta, P. (1997): Crack front propagation and fracture in a graphite sheet: a molecular - dynamics study on parallel computers. *Physical Review Letters*, vol. 78, pp. 2148 - 2151