DOI: 10.1002/chem.200601255

The Use of Quantum Molecular Calculations to Guide a Genetic Algorithm: A Way to Search for New Chemistry

Marcus C. Durrant*[a]

Abstract: The process of gene-based molecular evolution has been simulated in silico by using massively parallel density functional theory quantum calculations, coupled with a genetic algorithm, to test for fitness with respect to a target chemical reaction in populations of genetically encoded molecules. The goal of this study was the identification of transition-metal complexes capable of mediating a known reaction, namely the cleavage of N_2 to give the metal nitride. Each complex within the search space was uniquely specified by a nanogene consisting of an eight-digit number. Propagation of an individual nanogene into successive generations was determined by the fitness of its phenotypic molecule to perform the target reaction and new generations were created by recombination and mutation of surviving nanogenes. In its simplest implementation, the quantum-directed genetic algorithm (QDGA) quickly located a local minimum on the evolutionary fitness hypersurface, but proved incapable of progressing towards the global minimum. A strategy for progressing beyond local minima consistent with the Darwinian paradigm by the use of environmental variations coupled with mass extinctions was therefore developed. This allowed for the identification of nitriding complexes that are very closely related to known examples from the chemical literature. Examples of mutations that appear to be beneficial at the genetic level but prove to be harmful at the phenotypic level are described. As well as revealing fundamental aspects of molecular evolution, QDGA appears to be a powerful tool for the identification of lead compounds capable of carrying out a target chemical reaction.

Keywords: density functional calculations · genetic algorithms · molecular evolution · nitrides · reaction mechanisms

Introduction

The ultimate basis of natural selection depends on the efficiencies with which individual molecular species within an organism perform their associated chemical reactions. The genome of an organism contains sufficient information to synthesise a vast range of molecules, but for many species, such as metabolites, the information is indirectly coded and must be elaborated through assemblies of enzymes. This suggests that virtually any type of molecule can be subject to genetic evolution. Computational modelling of the general process of evolution has led to the development of genetic algorithms (GAs).[1] As well as providing valuable insights into biological evolution, GAs have found use in a very diverse range of applications. Nevertheless, at present there are relatively few examples of GAs that have been applied to chemical problems. The possibility of combining GA approaches with the calculation of molecular properties is therefore intriguing in terms of both the more realistic modelling of evolutionary processes and the development of new chemistry. Several research groups[2] have recently demonstrated the power of computational methods in the development of new protein functionalities, publishing papers that describe the use of semiempirical potential functions of various forms to calculate protein–ligand affinities, whereas Goldstein and co-workers have studied fundamental aspects of protein evolution using simple lattice models and contact energies to calculate thermodynamic properties.[3] In order to progress to studies of the evolution of chemical reactivity, however, it would be necessary to resort to quantum calculations and proteins are at present far too large for the routine application of quantum methods such as density functional theory (DFT). Recently, Jóhannesson et al. successfully used a combination of DFT calculations and an evolutionary approach to search for stable four-component alloys

[a] Dr. M. C. Durrant
Biomolecular and Biomedical Research Centre
School of Applied Sciences, Ellison Building
Northumbria University
Newcastle upon Tyne, NE1 8ST (UK)
Fax: (+44) 191-227-3519
E-mail: marcus.durrant@unn.ac.uk

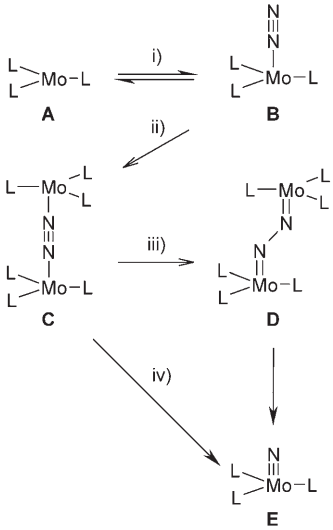
Supporting information for this article is available on the WWW under http://www.chemeurj.org/ or from the author.

based on the calculation of heats of formation.[4] For suitably small molecules, it should also be possible to calculate reaction profiles for populations of genetically encoded molecules, and by so doing, to simulate the process of genetic molecular evolution towards specific chemical reactivities. In this regard, transition-metal complexes appear to be suitable candidates in that they have a very rich chemistry and yet their three-dimensional structures can be defined relatively easily in terms of short numerical strings (nanogenes). This study was designed to answer the following questions: 1) Can the evolution of a population of molecules towards a desired chemical reactivity by selection of their associated nanogenes be demonstrated? 2) How do selection criteria based on the energetics of a target reaction drive the evolutionary process? 3) Can in silico simulation of molecular evolution be harnessed to develop new lead compounds for catalysis?

For this initial study, a simple chemical reaction, rather than a catalytic cycle, was chosen to define the environment against which individuals would be selected. This target reaction is a nitriding process described by Cummins and coworkers [Equation (1), where Ar = 3,5-$Me_2C_6H_3$].[5]

$$2\,[\mathrm{Mo}\{\mathrm{N}(t\mathrm{Bu})\mathrm{Ar}\}_3] + \mathrm{N}_2 \rightarrow 2\,[\mathrm{Mo(N)}\{\mathrm{N}(t\mathrm{Bu})\mathrm{Ar}\}_3] \qquad (1)$$

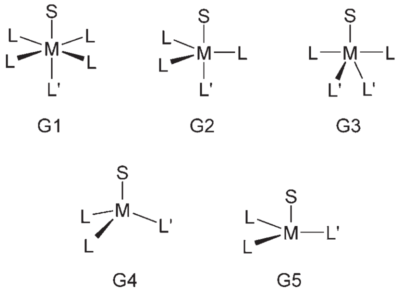
In addition to Cummins and co-workers' extensive experimental studies, the reaction has also been investigated in detail through DFT calculations.[6] Therefore, this reaction provides a body of experimental and theoretical data against which the results of the QDGA search could be evaluated. For the purposes of this study, it was necessary to break down Equation (1) into four elementary steps for which individual reaction energies can be calculated, as shown in Scheme 1. Initial capture of N_2 by **A** to give **B**, step i), is reversible; this is important because irreversible N_2 capture would mean that there would be insufficient free [Mo{N(*t*Bu)Ar}$_3$] in solution to allow the formation of the N_2-bridged dimer **C**. The dimer then passes through a transition state **D** to give two molecules of the product **E**. By considering the published data on Equation (1), we can assign "ideal" energies to each of the steps i)–iv); any complex whose calculated properties match these ideal values should then be capable of undergoing an analogous nitriding reaction.



Scheme 1. Elementary steps in the nitriding reaction of Cummins' complex, used to define the fitness functions used in this work. The equivalent reaction energies were calculated for individual complexes coded by nanogenes in this study. L denotes the ligand {N(*t*Bu)Ar}.

Computational Methods

The metal complexes to be tested were specified by nanogenes consisting of eight-digit numbers, as follows. The first two digits specify the row and column of the transition metal, for example, 10, 23 and 38 specify scandium, molybdenum and silver, respectively. The 27 available transition metals thus require allowed values of 1–3 for the first digit and 0–8 for the second. The third digit specifies the molecular charge and takes allowed values of 9, 0 and 1, corresponding to charges of −1, 0 and +1, respectively. The fourth digit gives the coordination geometry around the metal (see Scheme 2). The fifth and sixth digits specify the primary ligand and the seventh and eighth digits specify the secondary ligand (see Scheme 2 caption). The total number of possible nanogenes using this system is then 27×3×5×10×14 = 56700. For each nanogene of interest, the reaction energies defined in Scheme 1 were calculated by means of individual DFT calculations on the species equivalent to **A**–**E** using the B3LYP functional and LanL2DZ basis set, as implemented in Gaussian 98[7a] running on standalone PCs and Gaussian 03[7b] on the UK HPCx national supercomputing facility (both versions of the program gave comparable energy values). For the mononuclear species, separate calculations were carried out on all electron spin states with five or less unpaired electrons. Initially, dinuclear species with up to 10 unpaired electrons were considered. However, because in the initial calculation sets the highest spin state was never the ground state, the highest spin-state calculations were subsequently omitted. The very large numbers of DFT calculations required for this study meant that of necessity, accuracy had to be sacrificed in favour of high throughput. Therefore, ligands were restricted to simple hydrides (OH_2, PH_3, SH^-, etc.) and geometry optimisations were run using loose convergence criteria only; the energies of the transition-state structure **D** were approximated by running geometry optimisations with a fixed N–N distance of 1.6 Å (based on the value found for the [Mo(NH_2)$_3$] model of Cummins' compound) rather than full saddle-point calculations. The geometries of the ligands were optimised by using z-matrix procedures, but the overall coordination geometries of the metal sites were kept fixed as octahedral, trigonal bipyramidal, tetra-



Scheme 2. Allowed metal coordination geometries used in this study; primary and secondary ligands are denoted by L and L′, respectively. Primary ligands were specified by the following codes: 11, NH_2^-; 12, OH^-; 13, F^-; 14, Cl^-; 15, PH_2^-; 16, SH^-; 21, NH_3; 22, OH_2; 23, PH_3; 24, SH_2. Secondary ligands were specified by the following codes: 41, NH_2^-; 42, OH^-; 43, F^-; 44, Cl^-; 45, PH_2^-; 46, SH^-; 47, cyclopentadienyl; 48, H^-; 51, NH_3; 52, OH_2; 53, PH_3; 54, SH_2; 62, O^{2-}; 64, S^{2-}.

hedral or trigonal; the rationale for this constraint was that in terms of chemical synthesis, the desired geometry could be imposed on a metal by the use of, for example, steric effects and chelating ligands. Note that fixing the geometry of the metal centre in this way can have a significant effect on the reaction energetics; for example, the [$Mo(NH_2)_3$] complex itself adopts a trigonal geometry, but the corresponding nitride [$Mo(N)(NH_2)_3$] prefers a near-tetrahedral configuration,[6] and the destabilisation resulting from the geometry constraint is almost 30 kcal mol^{-1} in the latter case. Nevertheless, the use of simple hydride ligands is likely to significantly underestimate steric effects compared with experimentally employed ligands such as N(*t*Bu)Ar and allow too much flexibility compared with experiment, offsetting some of this discrepancy. Although the choice of method employed meant that energies obtained from these calculations are fairly approximate, they appear to be sufficiently accurate for the purpose (see below). This is probably because the results were used on a comparative basis, so systematic errors, such as those resulting from the use of rigid metal geometries, tend to cancel out. As an additional means of increasing throughput, the initial phases of the investigation were restricted to partial data sets using only the mononuclear species **A**, **B** and **E** in Scheme 1, as described below. The total cost of the project was about 70000 CPU hours. Calculated reaction energies for all the nanogenes considered in this study are given in the Supporting Information.

Results and Discussion

In order to gain a preliminary insight into the feasibility of the QDGA method, an initial run of three generations was used to search for nitride complexes with strong metal–nitrogen bonds. The only species calculated for each nanogene were the starting complex **A** and its nitride **E**; in combination with the calculated energy of the free nitrogen atom, these were sufficient to give the metal–nitride bond dissociation energies (MN-BDEs). An initial population of 18 completely random nanogenes was created; these gave calculated MN-BDEs ranging from +6 to +156 kcal mol^{-1}. For convenience, the evolutionary fitness parameter *F* for this phase of the calculations was taken as the difference between the BDE of the N_2 molecule (calculated value +181 kcal mol^{-1}) and the MN-BDE. The fittest six nanogenes were then selected by simply choosing those with the lowest values of *F*. Generation 2 consisted of these six parents, plus six offspring, generated by random recombinations of the parents, and six mutants. Throughout this study, recombination and mutation processes were kept separate, that is, new nanogenes were the product of either process, but not of both. Recombinations were carried out as follows. For generations 1–3, the parents' nanogenes were pooled and new nanogenes were created by drawing codes (metal, charge, geometry or ligand) from the pool at random. For subsequent generations, a pairwise breeding scheme was implemented in which each nanogene in turn was assigned a breeding partner from the other survivors and a new nanogene created by drawing codes at random from the two parents. For mutations, one of the five parameters coded by the parent nanogene was randomly selected and modified. This was carried out by using the principle of minimisation of chemical change upon mutation; thus, only incremental changes in the molecular charge (by +/−1) were allowed, whilst mutations in the metal were restricted to replacement by one of its immediate neighbours in the periodic table. Generation 2 was then subjected to quantum calculations as before, six new survivors were selected and the whole process was repeated a third time. The results, shown graphically in Figure 1, show a clear progression towards a fitter population with each generation. The six fittest nanogenes from generation 3 had MN-BDEs in the range of +156 to +176 kcal mol^{-1}.

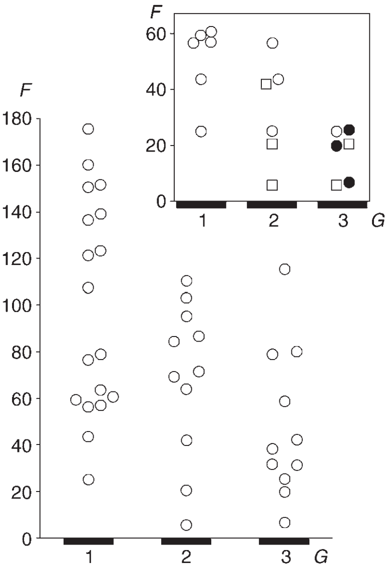


Figure 1. Evolution of strongly nitriding nanogenes in generations 1–3. The vertical axis is the fitness function *F*, defined here as *F* = 181−MN-BDE, and the horizontal axis is the generation number *G*. The main panel shows all the new complexes in each generation and the insert shows the survivors carried through into the next generation (complexes arising in generations 1, 2 and 3 are denoted by open circles, open squares and filled circles, respectively).

Encouraged by this preliminary result, I ran a further set of five generations, this time including all three monomeric species shown in Scheme 1. This allowed for refinement of the search to look for complexes with strong MN-BDEs, but relatively weak N_2 binding energies. It is generally accepted that new protein functionalities are acquired by exaptation from ancestral proteins with different chemical functions; this concept was applied in this study by considering that evolution towards the more complicated, multistep chemical requirements specified in Scheme 1 might be possible by starting from nanogenes bred for simpler, related reactivity patterns. If correct, this might allow for a more economical QDGA procedure overall as the CPU times required for calculations on the dimeric species **C** and **D** were far greater than those required for the monomers; typically ≈20 times longer.[8] During this part of the investigation, several problems were identified and a number of modifications to the procedure were introduced, as now described. First, the principle of minimisation of chemical change upon mutation, discussed above, was extended to the ligands through the introduction of allowable mutation matrices (full details of the mutation matrices and their use are given in the Supporting Information). Next, the use of more than one property to define the fitness function F requires a more refined approach. A general form of the equation for F is defined by Equation (2), where ΔE_n is the calculated energy for a particular elementary step and V_n is the assigned ideal energy for that step, based on the calculated energetics of the Cummins system.[6] As noted above, the Cummins reaction requires that the initial N_2 binding be reversible, hence both very strong and very weak N_2 binding should incur a penalty in the definition of F.[9] In contrast, the MN-BDE should be reasonably strong, but can never be too strong; to give an analogy, the top speed of a predator may be relatively unimportant provided that it is faster than its prey. In order to take this into account, several of the elementary steps considered in this work were assigned cut-off values such that if the calculated energy for the step was more favourable than that specified by the ideal value, this term of Equation (2) was set to zero. For this phase, the ideal values of the N_2 binding energy and MN-BDE were taken as −15 and +160 kcal mol^{-1}, respectively, and a cut-off was used for the MN-BDE term in the calculation of F. Survivors were drawn from all three previous generations, calculating the N_2 binding energies as required.

$$F^2 = \sum (\Delta E_n - V_n)^2 \quad (2)$$

Although generations 4–6 showed clear evolutionary progress in the form of a gradual improvement of the aggregate F values of the survivors, a notable feature of the survivors from generation 6 was that all of them had the same metal and primary ligand. This suggests that the small population size, dictated by the cost of the calculations, was leading to genetic inbreeding of the system. Since this would restrict the range of the evolutionary search within each new generation, a new condition was imposed on the selection of subsequent survivors. Under these enforced diversity selection rules, rather than choosing the six fittest members of the population, additional criteria were required. The nanogenes were ranked in order of increasing F and the six survivors were selected as follows: 1) best-ranked nanogene, that is, lowest F; 2) best-ranked nanogene with a different metal to (1); 3) best-ranked nanogene with a different geometry to (1); 4) best-ranked nanogene with a different primary ligand to (1); 5) best-ranked nanogene with a different secondary ligand to (1); 6) best-ranked nanogene with a different metal to those in both (1) and (2). This procedure was applied to the assembly of all nanogenes from generations 3–6, giving a new set of survivors from which to build generation 7. Generations 7 and 8 were then run as before.

The survivors from generation 8 showed only one new nanogene compared with their parents, suggesting that even with enforced diversity, this simplified system was close to its optimum solution. Therefore, at this point the selection procedures were changed to allow a direct search for the Cummins compound, including calculations on all of the species shown in Scheme 1. This required a new form of Equation (2) [Equation (3), where the subscripts 1–4 relate to the elementary reaction steps (i)–(iv) in Scheme 1].

$$F^2 = (\Delta E_1 + 10)^2 + (\Delta E_2 + 30)^2 + (\Delta E_3 - 30)^2 + (\Delta E_4)^2 \quad (3)$$

Hence, the following ideal energies were used: step i), −10 kcal mol^{-1} (reversible initial N_2 capture); step ii), −30 kcal mol^{-1} (irreversible dimer formation); step iii), +30 kcal mol^{-1} (barrier to dimer cleavage); step iv), 0 kcal mol^{-1} (dimer cleavage energetically neutral overall). Each of the last three terms in Equation (3) was set to zero in the event that $\Delta E_2 < -30$, $\Delta E_3 < +30$ or $\Delta E_4 < 0$ kcal mol^{-1}, respectively, following the logic discussed above. Generations 9 and 10 were then run using this extended fitness parameter and enforced diversity. Assessing the outputs from generation 10 suggested a difficulty in the current search strategy in that all six survivors from generation 10 fully satisfied the fitness tests for ΔE_3 and ΔE_4, and five of them also met the test for ΔE_2, but all six still gave quite poor results for the test for ΔE_1, with N_2 binding energies of −20 to −27 kcal mol^{-1}. In contrast, one of the discarded nanogenes from generation 10, namely 23142151, had a much better value of ΔE_1, −14 kcal mol^{-1}, but could not compete owing to its poor value of ΔE_2; −1 kcal mol^{-1} compared with a required value of −30 kcal mol^{-1} or lower. This suggests that the system was converging towards a local minimum and would be incapable of locating the Cummins compound. To explore this possibility further, a new function was introduced, the mutation step parameter M. This is defined as the number of individual mutations that would be required in order to convert any given nanogene into the nanogene for the Cummins compound, namely 23051141 (note that M may be greater than the Hamming distance owing to the restrictions on allowed mutations discussed above). The M values for the survivors of generation 10 were all between 5 and 7, whereas the value for the discarded nanogene 23142151 was 4, again suggesting that the search strategy in use would not be able to locate the Cummins compound. As a possible solution to this problem, a new set of calculations was set up using the concept of a niche environment. The niche calculations used a different fitness parameter, biased in favour of ΔE_1, given by Equation (4).

$$F_N^2 = (\Delta E_1 + 10)^2 + [(\Delta E_2 + 30)^2 + (\Delta E_3 - 30)^2 + (\Delta E_4)^2]/100 \quad (4)$$

To keep the number of calculations tractable, all the subsequent generations, both the main and niche [Eqs. (3) and (4), respectively], were reduced in size to five survivors plus 10 offspring and mutants. Three main and niche generations, denoted as generations 11M–13M and 11N–13N, respectively, were run in parallel. A final modification, introduced at this stage, was the use of a more drastic form of the enforced diversity algorithm; instead of selecting any different primary or secondary ligand as specified above, under the new scheme, the first digit of the two-digit codes specifying L1 and L2 was required to be different. This digit determines the charge on the ligand; anionic primary and secondary ligands were specified by L1 and L2 codes beginning with 1 and 4, respectively, whilst neutral ligands were specified by 2 and 5, respectively.

A notable feature of the outcome of this step was that the survivors from generation 13M were unchanged from generation 12M, suggesting that the procedure using the main fitness function, Equation (3), had indeed converged to a local minimum. In contrast, the niche population was still evolving, with two new survivors included in the output from generation 13N. Moreover, the aggregate M values for the survivors from generations 13M and 13N were 27 and 12, respectively, clearly showing that the niche population was more closely related to the Cummins compound than the main population at this stage. However, the aggregate F scores for the survivors from generations 13M and 13N were 79.0 and 131.3, respectively, moreover each individual niche survivor scored more poorly using F than its main generation counterpart. Hence, the niche population would not be able to compete successfully against the main population at this stage, even though it was now more closely related to the Cummins compound. In order to resolve this dilemma within a Darwinian framework, I assumed that the main population underwent a mass extinction at this point; the main population was abandoned and the niche population was used to repopulate the main environment. Thus, the calculations were continued for generations 14–21 by selecting the output from 13N and subsequent generations according to the main fitness parameter, Equation (3). The overall scheme for the entire QDGA procedure is summarised in Figure 2.

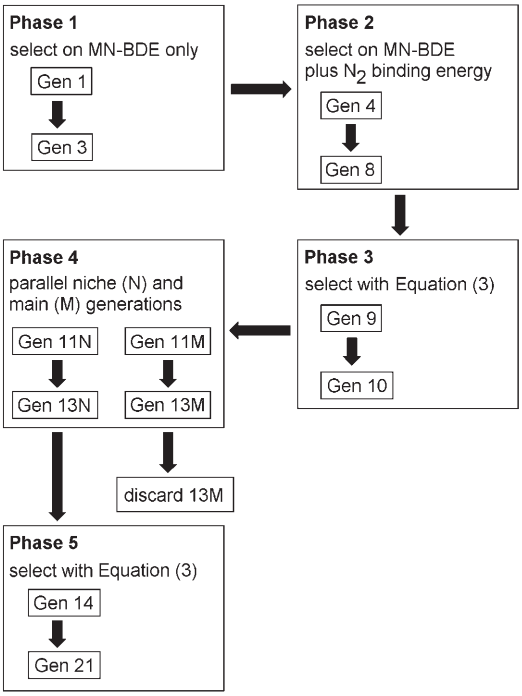


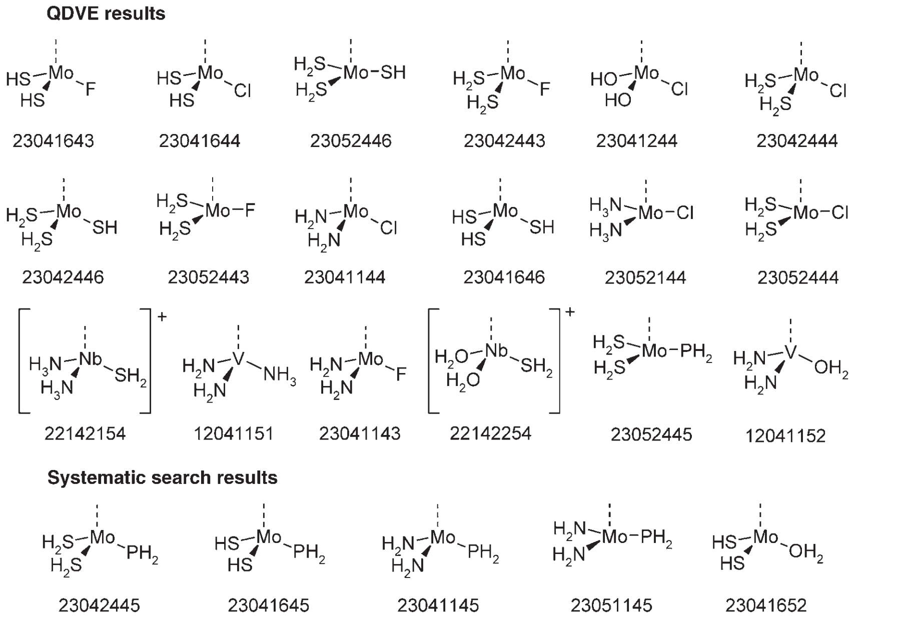
Figure 2. Summary of the overall QDGA procedure used in this study.

As the survivors of generation 21 were unchanged from their progenitors and also because all five survivors had F values of less than 10, the system was judged to have converged to an acceptable solution at this point. The total number of nanogenes, from all generations, for which $F<10$ was 18, as shown in the first section of Table 1, and the structures of the corresponding complexes are given in Scheme 3. Consideration of the results in Table 1 is instructive. First, all of these nanogenes code for three-coordinate complexes, with both the tetrahedral and trigonal geometries represented. Only three transition metals are included and the overall charge on the molecule correlates with the choice of metal; thus we find neutral Mo^{I} and Mo^{III}, neutral V^{II} (formally d^5, d^3 and d^3 complexes, respectively) and cationic Nb^{I} (d^4). The two niobium complexes are different from the others in that they originate from generation 12M (Figure 2) and so represent the closest approach of the main population to an optimum solution. Scheme 3 reveals quite wide variations in both primary and secondary ligands. Although there is no clear consensus, it is

Table 1. Reaction energies [kcal mol^{-1}] for all viable nanogenes from the QDGA and systematic searches.

Nanogene	Generation[a]	ΔE_1	ΔE_2	ΔE_3	ΔE_4	F
QDGA results						
23041643	20	−9.9	−46.5	+29.8	−4.0	0.1
23041644	19	−9.6	−45.0	+29.6	−3.0	0.4
23052446	15	−11.4	−32.6	+31.3	+0.4	2.0
23042443	20	−12.3	−43.6	+31.3	−16.4	2.6
23041244	18	−13.2	−39.2	+27.4	−10.2	3.2
23042444	19	−13.3	−34.0	+18.7	−21.0	3.3
23042446	18	−14.1	−42.7	+27.2	−14.5	4.1
23052443	21	−14.4	−29.7	+24.1	+6.4	7.8
23041144	20	−16.0	−47.5	+16.7	−14.3	6.0
23041646	21	−16.1	−38.6	+30.0	−12.7	6.1
23052144	15	−15.2	−26.4	+18.6	−2.1	6.3
23052444	19	−8.8	−33.4	+24.5	+6.5	6.6
22142154	12M	−17.8	−36.9	+24.4	−28.0	7.8
12041151	13N	−18.1	−45.4	+29.2	+1.0	8.2
23041143	21	−18.7	−48.0	+14.5	−18.2	8.7
22142254	12M	−18.7	−32.2	+27.4	−38.7	8.7
23052445	20	−19.2	−30.7	+29.2	+0.4	9.2
12041152	18	−17.9	−45.1	+26.4	+4.8	9.2
systematic search results						
23042445	–	−13.6	−38.4	+25.7	−14.1	3.6
23041645	–	−15.7	−45.3	+30.6	−1.5	5.7
23041145	–	−16.9	−54.5	+22.8	−11.2	6.9
23051145	–	−12.6	−34.1	+34.5	+7.4	9.1
23041652	–	−19.7	−37.1	+21.6	−2.4	9.7

[a] Generation in which the nanogene first appeared; M and N denote main and niche generations, respectively, see Figure 2.



Scheme 3. Structural formulae of the complexes corresponding to the nanogenes in Table 1. The dashes indicate the position of the substrate.

noteworthy that all of the molybdenum and vanadium complexes include at least one anionic ligand. This result can be compared with the calculations of Christian et al. who found that the energetics of dinitrogen cleavage by homoleptic three-coordinate molybdenum complexes are more favourable for complexes with anionic ligands than their neutral counterparts.[6a] The two niobium complexes in Scheme 3 are again different in that they only contain neutral ligands.

Table 2. Reaction energies [$\mathrm{kcal\,mol^{-1}}$] for nanogenes closely related to that of the Cummins compound.

Nanogene	Formula	ΔE_1	ΔE_2	ΔE_3	ΔE_4	F	M[a]
23051141	$[Mo(NH_2)_3]$	−10.2	−31.7	+35.4	+6.2	8.2	0
23052141	$[Mo(NH_3)_2(NH_2)]$	−27.8	−14.6	+24.7	+1.5	23.6	1
23051151	$[Mo(NH_2)_2(NH_3)]$	−34.8	−15.9	+40.8	+22.1	37.7	1
23052151	$[Mo(NH_3)_3]$	−20.2	−23.2	+12.5	−14.4	12.3	2
23151151	$[Mo(NH_2)_2(NH_3)]^+$	−10.3	+18.0	+40.7	−33.9	49.2	2
23141151	$[Mo(NH_2)_2(NH_3)]^+$	−9.9	−3.6	+18.2	−53.9	26.4	3
12941141	$[V(NH_2)_3]^-$	−32.7	+16.3	+36.7	−65.3	52.0	3
12951141	$[V(NH_2)_3]^-$	−16.2	+31.6	+51.8	−31.0	65.6	2
22951141	$[Nb(NH_2)_3]^-$	−30.9	+8.2	+20.2	−71.8	43.5	2

[a] Number of individual mutations required to convert the specified nanogene into 23051141.

An important observation from the evolutionary point of view is that similarity between two nanogenes is not necessarily reflected by similar reaction energetics for their complexes. This is illustrated by the properties of complexes derived from nanogenes closely related to that of the Cummins compound (see Table 2 and Scheme S1 in the Supporting Information). The tetrahedral and trigonal complexes $[V(NH_2)_3]^-$ and $[Nb(NH_2)_3]^-$ are isoelectronic with $[Mo(NH_2)_3]$, but their rather strong first N_2 binding energies ΔE_1 and their very weak second N_2 binding energies ΔE_2 make them a poor choice for the target reaction even though the overall reaction energies are favourable. The trigonal and tetrahedral forms of $[Mo(NH_2)_2(NH_3)]^+$ are also isoelectronic with $[Mo(NH_2)_3]$ and both have favourable first N_2 binding energies, but the second N_2 binding energies are again poor. Of particular interest are the complexes $[Mo(NH_3)_2(NH_2)]$ and $[Mo(NH_2)_2(NH_3)]$, which both provide a direct evolutionary link between $[Mo(NH_3)_3]$ and $[Mo(NH_2)_3]$. Although the nanogenes of the two mixed ligand complexes are superior to that of $[Mo(NH_3)_3]$ in terms of relatedness to $[Mo(NH_2)_3]$ (cf. M values in Table 2), both show significantly poorer fitness scores as a result of their strong initial N_2 binding energies. Hence, the evolutionary connection between $[Mo(NH_3)_3]$ and $[Mo(NH_2)_3]$ is much harder to find than might be supposed and mutations that appear beneficial at the genetic level may prove to be harmful at the molecular level. This observation agrees with recent studies on the mechanism of evolution of antibiotic resistance, which showed that most of the hypothetical evolutionary pathways connecting a non-resistant protein with a resistant homologue were inaccessible in practice and that individual amino acid mutations could be either beneficial or harmful depending on the order in which they occurred.[10] One may also conclude from the present work that the evolutionary fitness hypersurface varies relatively predictably with the choice of metal and overall geometry, but varies unpredictably and dramatically with choice of ligand. It is not yet possible to say whether these attributes are particular to the problem in hand or whether the same pattern would be observed in other systems. If this proved to be the case, it would have interesting implications for the evolution of metalloenzymes, where the choice of amino acids capable of acting as ligands to a transition metal is quite restricted.

How does the output from these QDGA calculations compare with known experimental chemistry? Although the calculations did not locate the nanogene corresponding to the Cummins complex itself, this is not particularly surprising given the random nature of the evolutionary process. In terms of the identification of lead compounds for chemical synthesis, not all of the nanogenes from the QDGA search would be realistic synthetic targets as they included halide ligands whose chemistry would be difficult to control. Therefore, as a final step, a limited systematic search was carried out using all permutations of the surviving non-halide-containing nanogenes coding for molybdenum and vanadium complexes. This identified a further five nanogenes for which $F<10$, including 23051145 (systematic search section of Table 1 and Scheme 3). This nanogene differs from that of the Cummins compound only in the replacement of one nitrogen ligand by its phosphorus analogue and would suggest the Cummins compound as an obvious choice in terms of ease of synthesis. Regarding the vanadium and niobium complexes located by the QDGA search, it is interesting to note that the only other well-characterised examples of N_2 bond cleavage reported to date involve molybdenum, vanadium and niobium,[11] although as expected, real world chemistry is generally more complicated.

Conclusion

The results of this study show that genetic molecular evolution can be simulated in silico by means of DFT calculations optimised for high throughput rather than accuracy. The combined QDGA search followed by a systematic search correctly identified Cummins-type complexes as potential mediators of the target reaction, suggesting that this approach would be valuable as a first step towards the development of new chemistry. The QDGA results also raise a number of interesting questions in terms of the actual mechanisms of molecular evolution in vivo. Perhaps the most striking observation is that many mutations, although apparently beneficial at the genetic level (in the sense that the mutant nanogenes more closely resemble a superior nanogene than their parents), prove to be harmful at the molecular level. This is illustrated by Figure 3, which is a plot of F versus generation number for the survivors of generations 9–20. As expected, the population gradually evolves to resemble the best found solution, nanogene 23041643. Nevertheless, the correlation between similarity to 23041643 and fitness is quite weak; several of the nanogenes in generations 9–18 are quite similar to 23041643, but show poor fitness and do not survive into subsequent generations. The use of the niche population was critical in establishing a population related closely to the best found solution, even though the final niche population was still less fit than its main population counterpart. Overall, the evolutionary process revealed by this study is not very well described by the popular analogy[12] of a combination lock; a better analogy might be offered by the Rubik's cube in that as the optimum solution is approached, it becomes harder to reach by incremental changes. This seems to result from the underlying chemistry rather than the properties of the GA. Nevertheless, many of the features commonly observed in other GAs are notable here also. For example, there is a tension between the severity of selection pressures and the ability of a population to enter new areas of evolutionary space; stringent selection pressures tend to keep the population fixed around a local minimum on the evolutionary fitness hypersurface and so prevent the emergence of new traits that may prove to be beneficial in later generations. The principle of minimisation of chemical change upon mutation, employed in this study, probably also improves the efficiency of progress towards local minima at the expense of the global minimum.

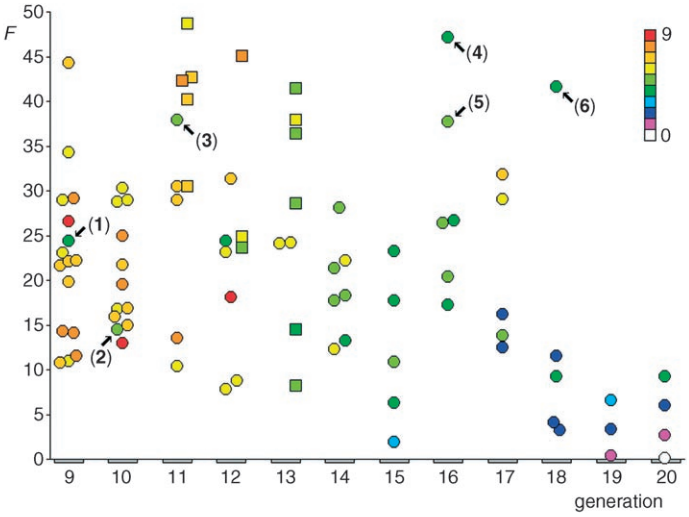


Figure 3. Plot of F versus the generation number for nanogenes from generations 9–20. The colour coding represents M', the similarity of each nanogene to the best found solution, 23041643, determined by the number of individual mutation steps required to convert the nanogene into 23041643. Circles and squares denote nanogenes from the main and niche populations, respectively. The arrows indicate nanogenes relatively closely related to 23041643, but not surviving into the next generation, as follows: 1) 22042142 ($M'=4$), 2) 22142452 ($M'=5$), 3) 23142152 ($M'=5$), 4) 12941146 ($M'=4$), 5) 23051151 ($M'=5$), and 6) 12941246 ($M'=4$).

Finally, it is interesting to consider that this study used reaction energies as the basis for selection, but in reality biomolecules must be selected based on equilibrium and rate constants, both of which are related to free energies by exponential functions. The range of values over which the position of a chemical equilibrium varies significantly with ΔG is only about 10 kcal mol^{-1}. Similar observations may be made for reaction rates; the energy difference between a process that occurs once per second and once per day is about 7 kcal mol^{-1}, comparable to the energy of a strong hydrogen bond. The variation in reaction energies for the elementary steps in Scheme 1 is at least an order of magnitude greater than this value (cf. Figure 1). Hence, a significant advantage of molecular evolution in silico is its ability to detect beneficial mutations in compounds whose ability to carry out a target reaction in vitro would be negligibly small. An obvious limitation of the theoretical approach is that it can only find what is looked for and so cannot detect unexpected side-reactions. However, this can arguably be seen as an advantage in that it can identify which systems are worth persevering in the laboratory. We are currently using Monte Carlo methods to investigate the effects of variations in the QDGA on the results obtained and employing QDGAs to search for transition-metal complexes capable of mediating catalytic cycles.

Acknowledgements

This work was supported through the BBSRC/IBM Life Science Initiative, grant reference BBSB13942. The help of staff at the Council for the Central Laboratory for the Research Councils and the Edinburgh Parallel Computing Centre in setting up calculations on HPCx is gratefully acknowledged.

[1] a) M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, **1996**; b) D. A. Coley, *An Introduction to Genetic Algorithms for Scientists and Engineers*, World Scientific, Singapore, **1999**.

[2] a) M. A. Dwyer, L. L. Looger, H. W. Hellinga, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11255–11260; b) M. A. Dwyer, L. L. Looger, H. W. Hellinga, *Science* **2004**, *304*, 1967–1971; c) M. Allert, S. S. Rizk, L. L. Looger, H. W. Hellinga, *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7907–7912; d) L. L. Looger, M. A. Dwyer, J. J. Smith, H. W. Hellinga, *Nature* **2003**, *423*, 185–190; e) J. K. Lassila, J. R. Keefe, P. Oelschlaeger, S. L. Mayo, *Protein Eng. Des. Sel.* **2005**, *18*, 161–163; f) D. N. Bolon, C. A. Voigt, S. L. Mayo, *Curr. Opin. Chem. Biol.* **2002**, *6*, 125–129; g) C. A. Voigt, S. L. Mayo, F. H. Arnold, Z.-G. Wang, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3778–3783; h) D. N. Bolon, S. L. Mayo, *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14274–14279; i) V. D. Sood, D. Baker, *J. Mol. Biol.* **2006**, *357*, 917–927; j) T. Kortemme, D. Baker, *Curr. Opin. Chem. Biol.* **2004**, *8*, 91–97; k) S. Ventura, L. Serrano, *Proteins Struct. Funct. Bioinf.* **2004**, *56*, 1–10.

[3] a) S. Govindarajan, R. A. Goldstein, *Proteins* **1997**, *29*, 461–466; b) D. M. Taverna, R. A. Goldstein, *Biopolymers* **2000**, *53*, 1–8; c) P. D. Williams, D. D. Pollock, R. A. Goldstein, *J. Mol. Graphics Modell.* **2001**, *19*, 150–156.

[4] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, J. K. Nørskov, *Phys. Rev. Lett.* **2002**, *88*, 255506.

[5] a) C. E. Laplaza, C. C. Cummins, *Science* **1995**, *268*, 861–863; b) C. E. Laplaza, M. J. A. Johnson, J. C. Peters, A. L. Odom, E. Kim, C. C. Cummins, G. N. George, I. J. Pickering, *J. Am. Chem. Soc.* **1996**, *118*, 8623–8638; c) C. C. Cummins, *Chem. Commun.* **1998**, 1777–1786.

[6] a) G. Christian, J. Driver, R. Stranger, *Faraday Discuss.* **2003**, *124*, 331–341; b) K. M. Neyman, V. A. Nasluzov, J. Hahn, C. R. Landis, N. Rosch, *Organometallics* **1997**, *16*, 995–1000; c) Q. Cui, D. G. Musaev, M. Svensson, S. Sieber, K. Morokuma, *J. Am. Chem. Soc.* **1995**, *117*, 12366–12367; d) M. C. Durrant, *Inorg. Chem. Commun.* **2001**, *4*, 60–62.

[7] a) Gaussian 98 (Revision A.11), M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, N. Rega, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, J. A. Pople, Gaussian, Inc., Pittsburgh, PA, **2001**; b) Gaussian 03, Revision C.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople, Gaussian, Inc., Wallingford CT, **2004**.

[8] In order to probe the possibility of using exaptation in more detail and to test the effects of the other modifications to the original procedure, namely, enforced diversity and the use of allowed mutation matrices, an additional set of four generations was run, starting again from the original generation 1 population. Full details are given in the Supporting Information, but two features are worth noting here. First, when selection was based on the MN-BDE alone, there was a marginal deterioration in aggregate fitness as measured by the combined fitness parameter using MN-BDE plus the N_2 binding energy; gains in the first parameter were offset by deterioration in the second. Hence, the value of exaptation in improving the efficiency of QDGA procedures has not yet been demonstrated. Secondly, although the modified procedure gave nanogenes with a significantly improved aggregate fitness, careful examination of the ancestries of these nanogenes suggested that the improvements could be a result of fortuitous random choices rather than the modifications to the procedure. Many more runs would be required in order to build up a statistically valid view of the contributions from each modification.

[9] This point has not always been fully recognised in previous studies; for example, although the N_2 cleavage step is more favourable for $[W(NH_2)_3]$ than for $[Mo(NH_2)_3]$, the initial N_2 capture step is also significantly more exothermic for tungsten than for molybdenum. Estimation of the N_2 binding energies of these complexes is complicated by the fact that DFT calculations on $[M(NH_2)_3]$ species are very sensitive to the rotation of the NH_2 ligands about the M–N bond axes, but the preferred rotation would be prohibited by the steric bulk of the N(*t*Bu)(Ar) ligands in the experimental system. This has been allowed for in the present calculations by constraining all three NH_2 groups to keep the same torsion angle with respect to the threefold axis of the $[Mo(NH_2)_3]$ core, giving M–N_2 binding energies of -13.3 and -21.5 kcal mol^{-1} for the molybdenum and tungsten systems, respectively. In practice, then, the reaction of $[W(NH_2)_3]$ with N_2 would essentially go to completion, preventing formation of the dimer $[\{W(NH_2)_3\}_2(N_2)]$.

[10] D. M. Weinreich, N. F. Delaney, M. A. DePristo, D. L. Hartl, *Science* **2006**, *312*, 111–114.

[11] a) H. Kawaguchi, T. Matsuo, *Angew. Chem.* **2002**, *114*, 2916–2918; *Angew. Chem. Int. Ed.* **2002**, *41*, 2792–2794; b) M. D. Fryzuk, C. M. Kozak, M. R. Bowdridge, B. O. Patrick, S. J. Rettig, *J. Am. Chem. Soc.* **2002**, *124*, 8389–8397; c) G. K. B. Clentsmith, V. M. E. Bates, P. B. Hitchcock, F. G. N. Cloke, *J. Am. Chem. Soc.* **1999**, *121*, 10444–10445; d) A. Zanotti-Gerosa, E. Solari, L. Giannini, C. Floriani, A. Chiesi-Villa, C. Rizzoli, *J. Am. Chem. Soc.* **1998**, *120*, 437–438; e) E. Solari, C. Da Silva, B. Iacono, J. Hesschenbrouck, C. Rizzoli, R. Scopelliti, C. Floriani, *Angew. Chem.* **2001**, *113*, 4025–4027; *Angew. Chem. Int. Ed.* **2001**, *40*, 3907–3909; f) A. Caselli, E. Solari, R. Scopelliti, C. Floriani, N. Re, C. Rizzoli, A. Chiesi-Villa, *J. Am. Chem. Soc.* **2000**, *122*, 3652–3670; g) D. J. Mindiola, K. Meyer, J.-P. F. Cherry, T. A. Baker, C. C. Cummins, *Organometallics* **2000**, *19*, 1622–1624.

[12] R. Dawkins, *The Blind Watchmaker*, Longman, London, **1986**.

Received: September 6, 2006
Published online: January 16, 2007