Journal of Mathematical Chemistry 26 (1999) 87–94

Misgivings from the theoretical foundation of the path-following procedures based on the Elber–Karplus strategy

László L. Stachó [a], Gyula Dömötör [b] and Miklós I. Bán [b]

[a] *Bolyai Institute for Mathematics, JATE University of Szeged, Aradi Vértanúk tere 1, H-6720 Szeged, Hungary*
E-mail: stacho@math.u-szeged.hu

[b] *Institute of Physical Chemistry, JATE University of Szeged, P.O. Box 105, H-6701 Szeged, Hungary*
E-mail: domotor@chem.u-szeged.hu, m.i.ban@chem.u-szeged.hu

Received January 1999

A mathematical proof against the theoretical foundation of the Elber–Karplus (EK) global reaction path-following method and the improvements based on the EK strategy have been discussed. According to our arguments the minimization of the average value of the potential energy along a path to two energy minima never defines a reaction path (RP) unless in the chemically irrealistic situation where the points of the curve joining the two minima of reactants and products have constant energy values. Therefore, finding approximate RPs by EK-strategies for large chemical systems or even in mathematical test examples is impossible or at least strongly doubtful (the larger the system the more doubtful).

1. Introduction

In a former paper [33] – relied upon mathematical illustrations and test examples – we have criticised the original global path-following Elber–Karplus (EK) method [8] and the most important improvements [3,5,6,24,37] based on the EK strategy. Interfaced with molecular mechanics for calculating the energy of the system, one of these procedures [12] is in widely use especially for large (organic and bio-) molecules with a very large number of degrees of freedom. The actuality of our criticism is verified by the fact that an exact mathematical formulation of the methods using the EK strategy has still been missing and by the number of references [1,2,4,7,9,10,13–16,18,19,21–23,25,28,29,35,36,38–40] to the original EK method [8] that has not shown downward tendency even in the past few years. Nevertheless, it has been indicated [34] that the mathematical basis for such methods using minimization techniques is incorrect, therefore the results, even when they are in concordance with experimental data, should be accepted with reservations for simple chemical systems as well, and even for mathematical test functions [33]. Starting from the energy average of the line integral (or its discretized form) followed by minimizing this functional and applying penalty func-

tions as constraints lead in the EK-type calculations to the controversial results we have argued against. The method of Chiu et al. [3] is the first true improvement in the EK sequels. Although they are also starting from the line integral (or its discretized form) and use minimization, however, instead of using penalty functions they introduce a redistribution of the grid points to substitute the constraints employed in former versions of the EK method. Unfortunately, this redistribution is nothing else essentially than the homogenization procedure described in our earlier papers cited in [33], preceding Chiu's method by some years. This is the cause why even the author of an excellent general work about geometry optimization on potential energy surfaces [27] has been prone to stray by mistake into the trap of mixing the fundaments of the EK method (more precisely the Chiu's version [3]) with that of the DDRP method [30–32]. At this point we should also emphasize that, in contrast with tempting heuristical expectations, the discretized numerical implementation of the redistribution procedure may change the average energy along the path. This fact is the main reason of some practical success by redistribution, nevertheless, the algorithm does not produce a minimization of the energy avarage any longer. Therefore, the energy average minimization principle is (fortunately!) only a heuristical starting point for Chiu's method but by no means a mathematical foundation of it. When having set up our claims against the EK strategy in [33] we only gave the outlines of the exact mathematical proof of the arguments. However, we then indicated that the rigorous mathematical investigation of the problem and its evidences will be given in a next paper. The aim of this paper is to make up for this deficiency, by providing the detailed arguments promised.

2. Discussion

The reaction path (RP) of a given system of N atoms is a piecewise smooth curve $C:[0,1] \to \mathbb{R}^{3N}$ with arc length proportional parameterization in the coordinate configuration space $\mathbb{R}^{3N}$ connecting two local minima of the energy function $U:\mathbb{R}^{3N} \to \mathbb{R}$ in a manner such that its tangent vectors are parallel to the gradient of the energy function (in particular, the gradient vanishes at the breaking points of the curve C). By a minimum energy RP (MERP) we mean an RP C such that at each smooth point $C(s)$ of the curve C, the function U has a local minimum on the hyperplane

$$H_s := \left\{ p \in \mathbb{R}^{3N} \colon \left\langle p - C(s), \frac{\mathrm{d}}{\mathrm{d}s} C(s) \right\rangle = 0 \right\}$$

approaching orthogonally to the tangent vector of C through $C(s)$. If one's attention is restricted to sytems where the graph surface of the energy function U admits only one reaction valley and there is a unique RP which is a smooth MERP in the same time, several equivalent definitions [11,20] can be encountered in the literature. For the sake of unambiguity we are going to consider only such simple systems throughout

this work. In [8] a method has been described for finding the RP on the basis of the hypothesis that the RP minimizes the functional

$$\mathcal{A}: C \mapsto \frac{\int_0^1 U(C(s))\|\partial C/\partial s\| \, \mathrm{d}s}{\int_0^1 \|\partial C/\partial s\| \, \mathrm{d}s} \tag{1}$$

of the *energy average* (with respect to arc length proportional parameterization) for all smooth curves $C: [0, 1] \to \mathbb{R}^{3N}$ joining the two minima of the energy function U.

The primary mathematical aim of the present paper is to show that

$$\inf \mathcal{A} = \inf U$$

and this infimum is never attained if the energy function has only two local minima (at the configurations of reactants and products) as is the case in the simplest reactions. *The shape of a curve with low energy average values is in general very far from that of the RP.* Moreover, *if C is an piecewise smooth steepest descent path (SDP) in the configuration space starting from a minimum place of U such that the energy function U is not constant along C then C cannot be a local minimum of the average functional $\mathcal{A}$ in the sense of variational calculus* that is there exists a variational curve $\widetilde{C}$ such that the parameter value $t = 0$ is no local minimum place of the function $t \mapsto \mathcal{A}(C + t\widetilde{C})$.

Thus, unfortunately, *the mathematical foundation of the original EK method [8] for searching RP is false.* However, several examples have been cited in the literature where the EK method leads to quite correct approximating reaction curves for molecular systems with energy functions satisfying our hypothesis. What can be the reason for this paradox? Simple logical exclusion would suggest that this is possible only if the functional minimizing procedure applied to the average $\mathcal{A}$ in the numerical realization of the method is also false. For this purpose originally the Powell minimization algorithm [26] was applied. However, this algorithm has proved to be mathematically correct in finite dimensions (though there is no rigorous argument for its correct use with discretization in infinite dimensional manifolds of curves). This means that either the approximating discretized curves calculated by the EK method believed to be correct are incorrect and therefore the incorrect results have been misinterpreted, or the experimental and theoretical results coincide accidentally.

2.1. *Global minimization of the energy average functional*

Throughout this section U denotes a continuous real valued function $\mathbb{R}^K \to \mathbb{R}$ with $K > 1$ variables. Consider any two points p_1, p_2 of the space $\mathbb{R}^K$ and let $\varepsilon_1, \varepsilon_2, \ldots \downarrow \inf U$ be any sequence of real numbers converging strictly monotonically to the infimum (possibly the minimum) value of the function U. By the continuity of U, for every index n there exists an open ball B_n such that $U < \varepsilon_n$ on B_n. For each n, choose any two points q_{1n}, q_{2n} from B_n and let C_{1n} be a smooth curve with starting point p_1 and endpoint q_{1n}, furthermore, let C_{2n} be a smooth curve with starting point q_{2n} and endpoint p_2. Since balls are convex, given any two unit vectors $v_{1n}, v_{2n} \in \mathbb{R}^K$, the points q_{1n}, q_{2n} can be joined with smooth curves of arbitrarily large

length passing completely in B_n and having tangent vectors v_{1n}, v_{2n} at their starting and end points, respectively. In particular, for each n there is a curve G_n ranging in B_n such that

$$\operatorname{length}(G_n) > n\big(1 + \max|U(C_{1n})| + \max|U(C_{2n})|\big)\big(\operatorname{length}(C_{1n}) + \operatorname{length}(C_{2n})\big)$$

and the concatenated curve

$$C_n := C_{1n} \cup G_n \cup C_{2n}$$

is smooth (also at the points q_{1n} and q_{2n} of attachments). In general, an average value is the weighted arithmetic mean of the averages over a partition. Therefore, since $\mathcal{A}(G_n) < \varepsilon_n$ (because the curve G_n passes in B_n where the function U takes values $< \varepsilon_n$),

$$\begin{aligned}
\mathcal{A}(C_n) &= \frac{\operatorname{length}(C_{1n})}{\operatorname{length}(C_n)}\mathcal{A}(C_{1n}) + \frac{\operatorname{length}(G_n)}{\operatorname{length}(C_n)}\mathcal{A}(G_n) + \frac{\operatorname{length}(C_{2n})}{\operatorname{length}(C_n)}\mathcal{A}(C_{2n}) \\
&\leqslant \frac{\operatorname{length}(C_{1n})}{\operatorname{length}(G_n)}\mathcal{A}(C_{1n}) + \frac{\operatorname{length}(G_n)}{\operatorname{length}(G_n)}\mathcal{A}(G_n) + \frac{\operatorname{length}(C_{2n})}{\operatorname{length}(G_n)}\mathcal{A}(C_{2n}) \\
&\leqslant \frac{\operatorname{length}(C_{1n})}{n(1+\max|U(C_{1n})|)\cdot\operatorname{length}(C_{1n})}\mathcal{A}(C_{1n}) + \frac{\operatorname{length}(G_n)}{\operatorname{length}(G_n)}\mathcal{A}(G_n) \\
&\quad + \frac{\operatorname{length}(C_{2n})}{n(1+\max|U(C_{2n})|)\cdot\operatorname{length}(C_{2n})}\mathcal{A}(C_{2n}) \\
&\leqslant \frac{\max U(C_{1n})}{n(1+\max|U(C_{1n})|)} + \varepsilon_n + \frac{\max U(C_{2n})}{n(1+\max|U(C_{2n})|)} \to \inf U
\end{aligned}$$

as $n \to \infty$.

In chemically relevant situations we often have a bounded continuous energy function $U\colon \mathbb{R}^{3N} \to \mathbb{R}$ with only two local minima p_1 and p_2 where $U(p_2) = \min U < U(p_1)$. In this case the balls $B_n := \{x \in \mathbb{R}^{3N}\colon \|x - p_2\| < 1/n\}$ with the constants $\varepsilon_n := \sup U(B_n)$ suit our requirements. In this case it is enough to join p_1 and p_2 by any smooth curve C and continue C smoothly with a curve G_n of length $> n(1 + \max|U(C)|)$ passing in B_n and starting and ending in the point p_2. Then $\mathcal{A}(C \cup G_n) \to U(p_2) = \min U$. On the other hand, since $U(p_1) > U(p_2)$, for any rectifiable curve G joining p_1 with p_2 we have necessarily $\mathcal{A}(G) > U(p_2) = \min U$.

Example. Let $U : \mathbb{R}^{3N} \to \mathbb{R}$ be a continuous energy function having only two places of local minimum p_1, p_2 such that $U(p_1) > U(p_2) = \min U$. Let $C : [0, 1/2] \to \mathbb{R}^{3N}$ be the straight line segment passing from p_1 to p_2 and for any index n let $G_n : [1/2, 1] \to \mathbb{R}^{3N}$ be a circle tangent to the segment C with radius $1/n$ such that $G_n(1/2) = G_n(1) = p_2$. Then for the curves $C_n := C \cup G_1 \cup G_2 \cup \cdots \cup G_n$ we have $\lim_{n\to\infty} A(C_n) = U(p_2) = \min U$ while necessarily $\mathcal{A}(C) > U(p_2)$.

2.2. *The energy average from the view point of variational calculus*

Throughout this section we assume that the function $U:\mathbb{R}^K \to \mathbb{R}$ is continuously differentiable twice and $C:[0,1] \to \mathbb{R}^K$ denotes a *piecewise C^2-smooth stationary curve of the average functional $\mathcal{A}$ such that $U(C(0)) = \min U$ which is a SDP for U in the same time.* We assume without loss of generality that the curve C is of length 1 and is parameterized arc length proportionally, that is

$$1 = \operatorname{length}(C) = \left\| \frac{\partial C(s)}{\partial s} \right\| = \left\langle \frac{\partial C(s)}{\partial s}, \frac{\partial C(s)}{\partial s} \right\rangle^{1/2}. \tag{2}$$

The $\mathcal{A}$-stationary property of C is formulated as

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \mathcal{A}\big(C + t\widetilde{C}\big) = \left.\frac{\partial}{\partial t}\right|_{t=0} \frac{\int_0^1 U(C(s) + t\widetilde{C}(s)) \|\partial(C(s) + t\widetilde{C}(s))/\partial s\| \,\mathrm{d}s}{\int_0^1 \|\partial(C(s) + t\widetilde{C}(s))/\partial s\| \,\mathrm{d}s} = 0$$

for all smooth perturbations $\widetilde{C}:[0,1] \to \mathbb{R}^K$ with $\widetilde{C}(0) = \widetilde{C}(1) = 0$. This means that

$$\begin{aligned} &\left.\frac{\partial}{\partial t}\right|_{t=0} \int_0^1 U\big(C(s) + t\widetilde{C}(s)\big) \left\| \frac{\partial(C(s) + t\widetilde{C}(s))}{\partial s} \right\| \mathrm{d}s \\ &\quad = \int_0^1 U\big(C(s)\big) \left\| \frac{\partial C(s)}{\partial s} \right\| \mathrm{d}s \cdot \left.\frac{\partial}{\partial t}\right|_{t=0} \int_0^1 \left\| \frac{\partial(C(s) + t\widetilde{C}(s))}{\partial s} \right\| \mathrm{d}s. \end{aligned}$$

Since $\|\partial C(s)/\partial s\| \equiv \operatorname{length}(C) = 1$ by assumption (2) and

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \left\| \frac{\partial(C(s) + t\widetilde{C}(s))}{\partial s} \right\| = \left\langle \frac{\partial}{\partial s} C(s), \frac{\partial}{\partial s} \widetilde{C}(s) \right\rangle \left\langle \frac{\partial}{\partial s} C(s), \frac{\partial}{\partial s} C(s) \right\rangle^{-1/2},$$

then denoting the gradient of U by ∇U, we have

$$\begin{aligned} &\int_0^1 \big\langle \nabla U\big(C(s)\big), \widetilde{C}(s) \big\rangle \mathrm{d}s + \int_0^1 U\big(C(s)\big) \left\langle \frac{\partial}{\partial s} C(s), \frac{\partial}{\partial s} \widetilde{C}(s) \right\rangle \mathrm{d}s \\ &\quad = \int_0^1 U\big(C(s)\big) \mathrm{d}s \int_0^1 \left\langle \frac{\partial}{\partial s} C(s), \frac{\partial}{\partial s} \widetilde{C}(s) \right\rangle \mathrm{d}s. \end{aligned}$$

Using partial integration we get

$$\begin{aligned} &\int_0^1 \big\langle \nabla U\big(C(s)\big), \widetilde{C}(s) \big\rangle \mathrm{d}s - \int_0^1 \left\langle \frac{\partial}{\partial s} \left[U\big(C(s)\big) \frac{\partial}{\partial s} C(s) \right], \widetilde{C}(s) \right\rangle \mathrm{d}s \\ &\quad = -\mathcal{A}(C) \int_0^1 \left\langle \frac{\partial^2}{\partial s^2} C(s), \widetilde{C}(s) \right\rangle \mathrm{d}s. \end{aligned}$$

By the Du Bois–Reymond's lemma [17],

$$-\nabla U\big(C(s)\big) + \frac{\partial}{\partial s} \left[U\big(C(s)\big) \frac{\partial}{\partial s} C(s) \right] = \mathcal{A}(C) \frac{\partial^2}{\partial s^2} C(s)$$

for all $s \in [0,1]$). Hence the Euler–Lagrange equation [17] characterizing the stationary curves of the functional $\mathcal{A}$ is

$$\nabla U\big(C(s)\big) = \left\langle \nabla U\big(C(s)\big), \frac{\partial}{\partial s}C(s)\right\rangle \frac{\partial}{\partial s}C(s) + \big[U\big(C(s)\big) - \mathcal{A}(C)\big]\frac{\partial^2}{\partial s^2}C(s). \quad (3)$$

Next we proceed to the question whether the arc-length parameterized $\mathcal{A}$-stationary SDP $C\colon [0,1] \to \mathbb{R}^K$ is such that

$$U(0) = \min U < \max_s U\big(C(s)\big)$$

can be a local minimum of $\mathcal{A}$ in the sense of variational calculus. By adding a suitable constant to the energy function, we assume without loss of generality also that

$$\mathcal{A}(C) = \int_0^1 U\big(C(s)\big)\,\mathrm{d}s = 0.$$

Since $\mu := \min U = U(C(0)) < \mathcal{A}(C) = 0$, there exists $a \in (0,1)$ such that

$$U\big(C(s)\big) < \mu/2 \quad \text{for } 0 \leqslant s \leqslant a.$$

Therefore equation (3) allows for C only to have vanishing second derivative on $[0,a]$. That is,

$$C(s) = C(0) + su \quad \text{if } 0 \leqslant s \leqslant a$$

for some unit vector $u \in \mathbb{R}^K$. Consider the perturbation

$$\widetilde{C}(s) := \begin{cases} -s^{1/2}(a-s)^{3/2}u & \text{for } 0 \leqslant s \leqslant a, \\ 0 & \text{for } a \leqslant s \leqslant 1. \end{cases}$$

This is a continuous vector-valued function $[0,1] \to \mathbb{R}^K$ having continuous derivative on the open interval $(0,1)$. For any fixed $t > 0$, we have $C(s) + t\widetilde{C}(s) = C(0) + [s - ts^{1/2}(a-s)^{3/2}]u$ and

$$\frac{\partial}{\partial s}\big[C + t\widetilde{C}\big] = \left[1 - \frac{ts^{-1/2}}{2}(a-s)^{1/2}(a-4s)\right]u \quad \text{if } 0 < s < a.$$

Thus the equation

$$\frac{\partial}{\partial s}\big[C + t\widetilde{C}\big] = 0$$

is equivalent to the polynomial equation of 3rd degree $(a-s)(a-4s)^2 - 4s/t^2 = 0$ which has a unique root s_t on $(0,a)$. Since

$$\lim_{s\downarrow 0}\left[1 - \frac{ts^{-1/2}}{2}(a-s)^{1/2}(a-4s)\right] = -\infty$$

and

$$\lim_{s\uparrow a}\left[1 - \frac{ts^{-1/2}}{2}(a-s)^{1/2}(a-4s)\right] = 1,$$

this means that the function $s - ts^{1/2}(a-s)^{3/2}$ has negative derivative on the interval $(0, s_t)$ and it has positive derivative on (s_t, a) and s_t is its minimum place with negative minimum value which we denote by $-\lambda_t$. Thus the curve $C + t\widetilde{C}$ passes first straight from the point $C(0)$ to $C(0) - \lambda_t u$ then it returns back on the same straight line to $C(0)$, thereafter it passes straight to the point $C(0) + au$ and finally it coincides with the rest of C. It follows (since $\mathcal{A}(C) = 0$)

$$\mathcal{A}\big(C + t\widetilde{C}\big) = \frac{2\int_0^{\lambda_t} U(C(0) - su)\,\mathrm{d}s + \int_0^1 U(C(s))\,\mathrm{d}s}{2\lambda_t + 1}$$
$$= \frac{2\int_0^{\lambda_t} U(C(0) - su)\,\mathrm{d}s}{2\lambda_t + 1} < 0 \tag{4}$$

whenever $t > 0$ is so small that the function U attains negative values on the straight line segment joining the points $C(0)$ and $C(0) - \lambda_t u$.

Acknowledgements

This work was supported by the Hungarian Scientific Research Fund OTKA, Grant No. T 020743. Many thanks are due to Profs. M. Karplus and R. Elber for having corrected an error in the manuscript.

References

[1] V.V. Bulatov, M. Nastar, J. Justo and S. Yip, Nucl. Instrum. Methods Phys. Res. B 121 (1997) 251.
[2] V.V. Bulatov, S. Yip and A.S. Argon, Philos. Magaz. A 72 (1995) 453.
[3] S.S.L. Chiu, J.J.W. McDouall and I.H. Hillier, J. Chem. Soc. Farad. Trans. 90 (1994) 1575.
[4] M.A. Collins, Adv. Chem. Phys. 93 (1996) 389.
[5] R. Czerminski and R. Elber, Int. J. Quantum Chem. 24 (1990) 167.
[6] R. Czerminski and R. Elber, J. Chem. Phys. 92 (1990) 5580.
[7] J.F. Diaz, B. Wroblowski and Y. Engelborghs, Biochem. 34 (1995) 12038.
[8] R. Elber and M. Karplus, Chem. Phys. Lett. 139 (1987) 375.
[9] M.A.E.C. Elkettani and J.C. Smith, Compt. Rend. Ser. III 319 (1996) 161.
[10] C. Guilbert, D. Perahia and L. Mouawad, Comput. Phys. Comm. 91 (1995) 263.
[11] D. Heidrich, W. Kliesch and W. Quapp, *Properties of Chemically Interesting Potential Energy Surfaces* (Springer, Berlin, 1991).
[12] S. Huston and J.W. Ponder, *TINKER: Software Tools for Molecular Design*, Version 3.5, October 1997 (Copyright Jay William Ponder, 1990–97).
[13] E. Jacoby, P. Kruger, J. Schlitter, D. Koper and A. Wollmer, Protein Engrg. 9 (1996) 113.
[14] F. Jensen, J. Chem. Phys. 102 (1995) 6706.
[15] R. Khare and M.E. Paulaitis, Chem. Engrg. Sci. 49 (1994) 2867.
[16] R. Khare and M.E. Paulaitis, Macromol. 28 (1995) 4495.
[17] G.A. Korn and Th.M. Korn, *Mathematical Handbook for Scientists and Enginers* (McGraw-Hill, New York, 1961).
[18] T. Lazaridis and M.E. Paulaitis, J. Am. Chem. Soc. 116 (1994) 1546.
[19] O. Marques and Y.H. Sanejouand, Proteins–Structure, Function and Genetics 23 (1995) 557.
[20] P.G. Mezey, *Potential Energy Hypersurfaces* (Elsevier, Amsterdam, 1987).

[21] L. Mouawad and D. Perahia, J. Mol. Biol. 258 (1996) 393.
[22] M. Nastar, V.V. Bulatov and S. Yip, Phys. Rev. B 53 (1996) 13521.
[23] I. Ohmine, J. Phys. Chem. 99 (1995) 6767.
[24] R. Olender and R. Elber, J. Mol. Struct. (Theochem) 398–399 (1997) 63.
[25] R. Poteau, F. Spiegelmann and P. Labastie, Z. Phys. D 30 (1994) 557.
[26] W.H. Press, B.P. Flannery S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1986) p. 294.
[27] H.B. Schlegel, in: *Modern Electronic Structure Theory*, Part I, Advanced Series in Physical Chemistry, Vol. 2, ed. D.R. Yarkony (World Scientific, Singapore, 1995) p. 482.
[28] J. Schlitter, M. Engels and P. Kruger, J. Mol. Graphics 12 (1994) 84.
[29] O.S. Smart, Chem. Phys. Lett. 222 (1994) 503.
[30] L.L. Stachó and M.I. Bán, J. Math. Chem. 11 (1992) 405.
[31] L.L. Stachó and M.I. Bán, Theor. Chim. Acta 83 (1992) 433.
[32] L.L. Stachó and M.I. Bán, Theor. Chim. Acta 84 (1993) 535.
[33] L.L. Stachó, Gy. Dömötör and M.I. Bán, Chem. Phys. Lett. (accepted for publication in 1999).
[34] L.L. Stachó, Gy. Dömötör, M.I. Bán and T. Csendes, J. Mol. Struct. (Theochem) 398–399 (1997) 111.
[35] J.E. Straub and J.K. Choi, J. Phys. Chem. 98 (1994) 10978.
[36] A. Thomas, M.J. Field and D. Perahia, J. Mol. Biol. 261 (1996) 490.
[37] A. Ulitsky and R. Elber, J. Chem. Phys. 92 (1990) 1510.
[38] A. Ulitsky and D. Shalloway, J. Chem. Phys. 106 (1997) 10099.
[39] J.C. Wang and K.A. Fichthorn, Langmuir 12 (1996) 139.
[40] R.P. Wang and K.A. Fichthorn, Phys. Rev. B 48 (1993) 18288.