

1.3. Optymalizacja geometrii cząsteczki

Poszukiwanie punktów stacjonarnych (krytycznych) funkcji stanowi niezwykle istotny problem w obliczeniowej chemii kwantowej. Sprowadza się on do dwóch istotnych zagadnień: poszukiwania przybliżonych (wariacyjnych) rozwiązań równania Schrödingera dla układów wieloelektronowych oraz lokalizacji i klasyfikacji punktów stacjonarnych (minimów i punktów siodłowych niższych rzędów, *vide infra*) na hiperpowierzchni energii potencjalnej.

W przypadku lokalizacji i klasyfikacji punktów stacjonarnych na hiperpowierzchni energii potencjalnej optymalizacji poddaje się energię elektronową cząsteczki E . W przybliżeniu Born-Oppenheimera każdy stan elektronowy cząsteczki posiada własną powierzchnię energii potencjalnej PES (ang. *Potential Energy Surface*), zdefiniowaną w f -wymiarowej przestrzeni, gdzie f jest liczbą stopni swobody danej molekuly (*vide infra*). W znaczeniu takim powierzchnia energii potencjalnej cząsteczki jest matematyczną zależnością między jej energią a parametrami geometrycznymi. Problem optymalizacji geometrii cząsteczki sprowadza się do wyznaczenia struktury równowagowej, czyli takich współrzędnych jądrowych, dla których na hiperpowierzchni energii potencjalnej występuje minimum lokalne. Stany przejściowe reakcji chemicznych odpowiadają punktom siodłowym na PES. Dla reakcji chemicznych możliwa jest także optymalizacja struktury układu reakcyjnego przy zadanej wartości tzw. współrzędnej reakcji, która mierzy postęp wzdłuż drogi energii minimalnej. Lokalizacja i klasyfikacja punktów stacjonarnych na powierzchni energii potencjalnej zostanie szerzej omówiona w niniejszym rozdziale.

1.3.1. Diagnostyka punktów stacjonarnych.

Energia potencjalna cząsteczki N -atomowej E , uwzględniająca wzajemne odpychanie jąder atomowych, zależy od tzw. współrzędnych wewnętrznych $\mathbf{Q} = (Q_1, Q_2, \dots, Q_f)$, gdzie $f = 3N - 6$ jest liczbą stopni swobody cząsteczki.¹⁴ Jako współrzędne wewnętrzne wybiera się na ogół długości wiązań, kąty walencyjne i torsyjne [19]. Punkty stacjonarne na hiperpowierzchni energii $E(\mathbf{Q})$ wyznacza się z warunku koniecznego istnienia ekstremum, tj.

$$\frac{\partial E}{\partial Q_i} = 0, \quad \text{dla } i = 1, 2, \dots, f. \quad (1.37)$$

¹⁴W przypadku cząsteczki liniowej $f = 3N - 5$.

Elementy $g_i(\mathbf{Q}) = \frac{\partial E}{\partial Q_i}$ tworzą wektor zwany gradientem $\mathbf{g} = (g_1, g_2, \dots, g_f)^{15}$. Dla zadanej geometrii $\mathbf{Q}$, gradient wyznacza ujemne siły działające na jądra atomowe. Siły te wskazują kierunek relaksacji struktury cząsteczki do położenia równowagowego $\mathbf{Q}^0$ (dla geometrii równowagowej $\mathbf{Q}^0$ siły te zanikają). Zależność (1.37) pozwala jedynie na wyznaczenie ekstremum lokalnego w punkcie $\mathbf{Q}^0$, czyli lokalizację punktów stacjonarnych na powierzchni energii potencjalnej. Natomiast podstawą dalszej ich diagnostyki jest wyznaczenie tensora drugich pochodnych cząstkowych energii elektronowej względem współrzędnych jądrowych – hessianu $\mathbf{H}(\mathbf{Q})$, tj.

$$\mathbf{H}(\mathbf{Q}) = \begin{bmatrix} \frac{\partial^2 E}{\partial^2 Q_1} & \frac{\partial^2 E}{\partial Q_1 \partial Q_2} & \dots & \frac{\partial^2 E}{\partial Q_1 \partial Q_f} \\ \frac{\partial^2 E}{\partial Q_2 \partial Q_1} & \frac{\partial^2 E}{\partial^2 Q_2} & \dots & \frac{\partial^2 E}{\partial Q_2 \partial Q_f} \\ \vdots & \vdots & \dots & \vdots \\ \frac{\partial^2 E}{\partial Q_f \partial Q_1} & \frac{\partial^2 E}{\partial Q_f \partial Q_2} & \dots & \frac{\partial^2 E}{\partial^2 Q_f} \end{bmatrix}. \quad (1.38)$$

Hessian $\mathbf{H}(\mathbf{Q})$, podobnie jak gradient $\mathbf{g}(\mathbf{Q})$, ma istotny sens fizyczny – określa bowiem stałe siłowe dla wychyleń jąder w otoczeniu danej geometrii układu $\mathbf{Q}$. Na podstawie znaków minorów wyznacznika $\det \mathbf{H}$ obliczonego dla geometrii równowagowej $\mathbf{Q}^0$ wyznaczonej z warunku (1.37) dokonuje się klasyfikacji punktów stacjonarnych na powierzchni energii elektronowej. Stosowne informacje na ten temat można znaleźć w typowych podręcznikach analizy matematycznej (patrz np. [20, 21]).

Hessian stanowi symetryczną macierz wymiaru $f \times f$ o f^2 elementach. Liczba niezerowych elementów macierzy $\mathbf{H}(\mathbf{Q})$ może ulec redukcji przez diagonalizację hessianu, która sprowadza się do rozwiązania zagadnienia własnego, tj.

$$\mathbf{H}\mathbf{O} = \mathbf{O}\mathbf{H}', \quad (1.39)$$

gdzie $\mathbf{O}$ jest macierzą ortogonalną, tj. $\mathbf{O}\mathbf{O}^T = \mathbf{O}^T\mathbf{O} = \mathbf{I}$ ($\mathbf{I}$ oznacza tu macierz jednostkową) natomiast $H'_{ij} = H_i \delta_{ij}$. Dzięki tej transformacji otrzymuje się macierz diagonalną $\mathbf{H}'$.

Diagonalizacja tensora kwadratowych stałych siłowych za pomocą transformacji ortogonalnej jest możliwa z uwagi na symetryczność macierzy $\mathbf{H}$. Odpowiada ona pewnemu obrotowi układu współrzędnych [1, 22]. Na przykład w układzie $x_1 x_2$ dla powierzchni dwuwymiarowej $E(\mathbf{x})$, gdzie $\mathbf{x} = (x_1, x_2)$, która może być lokalnie przybliżona przez elipsoidalne zagłębienie (w przypadku minimum) diagonalizacja polega na takim doborze układu współrzędnych $x'_1 x'_2$, aby osie elipsy pokrywały się z osiami układu primowanego. W ogólności współrzędne punktu w układzie primowanym $x'_1 x'_2$ powiązane są ze współrzędnymi punktu w układzie $x_1 x_2$

¹⁵Od tej pory przyjmuje się, że współrzędne wektora tworzą wektor wierszowy.

zależnością

$$x'_j = \sum_k x_k O_{kj}, \quad (1.40)$$

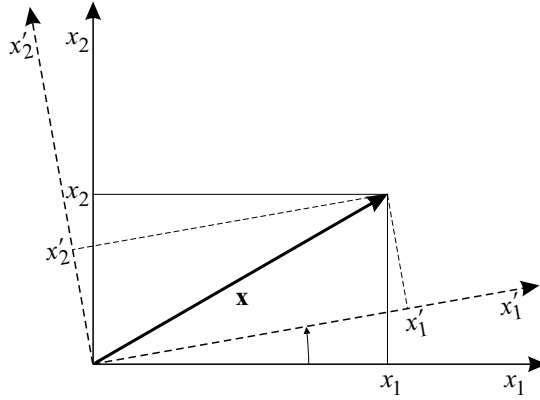
czyli

$$\mathbf{x}' = \mathbf{x}\mathbf{O}. \quad (1.41)$$

Ortogonalna macierz $\mathbf{O}$ dana jest jako

$$\mathbf{O} = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix}, \quad (1.42)$$

gdzie α oznacza kąt obrotu (patrz Rys. 1.2). Z prostego rachunku natomiast wynika, że



Rys. 1.2. Obrót układu współrzędnych o kąt α

$$g'_i = \frac{\partial E}{\partial x'_i} = \sum_j \frac{\partial E}{\partial x_j} \frac{\partial x_j}{\partial x'_i} = \sum_j \frac{\partial E}{\partial x_j} O_{ij}^T = \sum_j g_j O_{ji}, \quad (1.43)$$

czyli

$$\mathbf{g}' = \mathbf{g}\mathbf{O}, \quad (1.44)$$

gdzie odpowiednio $\mathbf{g}'$ i $\mathbf{g}$ są gradientami w układach $x'_1 x'_2$ i $x_1 x_2$. Macierz drugich pochodnych w układzie primowanym $\mathbf{H}'$ natomiast, po zróżniczkowaniu równania (1.43), dana jest jako

$$H'_{ij} = \sum_{k,l} O_{ik}^T H_{kl} O_{lj}, \quad (1.45)$$

czyli

$$\mathbf{H}' = \mathbf{O}^T \mathbf{H} \mathbf{O}. \quad (1.46)$$

Korzystając z własności macierzy ortogonalnej $\mathbf{O}$ zależność (1.46) przekształca się do równania (1.39), będącego zagadnieniem własnym hessianu. Jak więc widać, można wybrać taki

układ współrzędnych, w którym hessian będzie macierzą diagonalną. Powyższe rozważanie można z łatwością uogólnić na przestrzeń f -wymiarową. Odpowiadające wektorom własnym (kolumny macierzy $\mathbf{O}$) wartości własne mogą być wykorzystane do badania lokalnej topologii powierzchni energii potencjalnej – klasyfikacji punktów stacjonarnych. Dla minimum lokalnego, określającego geometrię równowagową cząsteczki i odpowiadającego stabilnym konformerom molekuly, wszystkie wartości własne położone na diagonalu macierzy $\mathbf{H}'$ są dodatnie. Punkt siodłowy k -tego rzędu charakteryzuje obecność wszystkich dodatnich wartości własnych z wyjątkiem k wartości ujemnych. Punkty siodłowe pierwszego rzędu ($k=1$) wyznaczają kompleksy przejściowe reakcji chemicznych. Występowanie wyłącznie ujemnych wartości własnych wskazuje na istnienie maksimum lokalnego. Maksima lokalne podobnie jak punkty siodłowe wyższych rzędów nie korespondują z żadną określoną strukturą cząsteczki [22].

1.3.2. Strategie optymalizacji

Najczęstszym problemem występującym przy procedurze minimalizacji jest odnalezienie globalnego minimum energii. Na powierzchni energii potencjalnej może wystąpić szereg minimów lokalnych, gdzie zerują się pierwsze pochodne, lecz nie jest tam osiągana najmniejsza jej wartość. Algorytmy minimalizacji często odnajdują strukturę cząsteczki odpowiadającą minimum lokalnemu. Szukając minimum globalnego należy rozpocząć obliczenia z różnych geometrii początkowych, czyli z odpowiadającym im różnych punktów na powierzchni energii potencjalnej. Ciągłe nie ma jednak pewności, że odnalezionej w ten sposób geometrii odpowiada minimum globalne. Zadaniem algorytmów optymalizacyjnych jest określenie kierunku i odległości do minimum od punktu początkowego. Najprostsze algorytmy na przykład poszukiwanie liniowe (ang. *line search*) lokalizują minimum poprzez odnalezienie jedynie kierunku, w którym należy się przesunąć od punktu $\mathbf{x}^k$, aby malały pochodne funkcji $f(\mathbf{x})$. Ponieważ dla dowolnego wektora $\mathbf{z}$ obowiązuje

$$\left(\frac{df(\mathbf{x}^k + t\mathbf{z})}{dt} \right)_{t=0} = \mathbf{g}^k \mathbf{z}^T < 0,$$

musi istnieć $\tau > 0$ takie, że

$$f(\mathbf{x}^k + t\mathbf{z}) < f(\mathbf{x}^k),$$

dla $0 < t < \tau$. Kierunek spadku pochodnej funkcji jest więc wyznaczony przez wektor $\mathbf{z}$. Istnieje wiele strategii minimalizacji, do najważniejszych należą: metoda największego spadku, metoda sprzężonych gradientów oraz metoda Newtona-Raphsona.

Metoda największego spadku (ang. *steepest descent*) wykorzystuje w każdym kroku iteracyjnym algorytm poszukiwania liniowego. Za kierunek największego spadku funkcji przyjmuje się gradient funkcji w danym punkcie wzięty ze znakiem przeciwnym. Następnie w wyniku minimalizacji $f(\mathbf{x}^k - t\mathbf{g}^k)$ względem t w kierunku tym odszukiwany jest punkt o najniższej wartości funkcji. Procedura powtarza się aż do odnalezienia minimum.

Metoda sprzężonych gradientów (ang. *conjugate gradient*) zwana jest także metodą zmiennej metryki. Algorytm ten wykorzystuje pierwsze pochodne (gradient) do wyznaczenia optymalnego kierunku dla liniowego poszukiwania minimum generowanego z punktu startowego $\mathbf{x}^k$ zgodnie z rekurencyjną formułą

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha^k \mathbf{A}^k \mathbf{g}^k, \quad (1.47)$$

gdzie α^k jest czynnikiem skalującym, $\mathbf{A}^k$ macierzą symetryczną, której postać zależy od przyjętej metryki, a $\mathbf{g}^k$ jest gradientem funkcji [19]. Algorytm ten może być niestabilny, jeżeli punkt startowy jest daleki od minimum.

Natomiast Metoda Newtona-Raphsona oprócz informacji, które niesie obliczanie gradientu, wykorzystuje macierz drugich pochodnych energii – hessian. Z uwagi na większą efektywność, stabilność i szybkość zbieżności w porównaniu z metodami największego spadku i sprzężonych gradientów, (co można uzyskać kontrolując długość kroku przy użyciu procedur RFO (ang. *Rational Function Optimization*) oraz TRM (ang. *Trust Radius Model*)) algorytm ten zostanie omówiony szerzej.

1.3.3. Metoda Newtona-Raphsona

Dla przypadku jednowymiarowego ekspansja funkcji $f(x)$, aproksymującej krzywą energii potencjalnej E , w szereg Taylora wokół punktu x^k dana jest jako

$$f(h) = f(x^k) + g(x^k)h + \frac{1}{2}H(x^k)h^2 + \dots, \quad (1.48)$$

gdzie $h = x - x^k$, $g = \frac{\partial f}{\partial h}$, $H = \frac{\partial^2 f}{\partial h^2}$. Przybliżanie funkcji E funkcją liniową uwzględniającą jedynie dwa pierwsze człony rozwinięcia charakteryzuje tzw. model liniowy. Nieograniczoność tego modelu wynika z braku punktów stacjonarnych dla funkcji postaci $f(h) \approx f(x^k) + g(x^k)h$, nie zawierającej żadnych informacji o krzywiźnie funkcji aproksymowanej E . Dyskwalifikuje to model liniowy w procedurach lokalizacji punktów krytycznych.

Biorąc pod uwagę trzy człony szeregu Taylora uzyskuje się funkcję postaci trójmianu kwadratowego. Model kwadratowy (paraboliczny) stanowi dobre przybliżenie powierzchni energii

potencjalnej w sąsiedztwie punktu x^k , w szczególności wokół jej punktów stacjonarnych. Z warunku koniecznego istnienia ekstremum, tj. $\frac{\partial f}{\partial h} = 0$, wynika, że

$$h = -\frac{g(x^k)}{H(x^k)}. \quad (1.49)$$

Równanie (1.49) stanowi podstawę metody Newtona-Raphsona dla jednego wymiaru, a wartość h nosi nazwę iteracyjnego kroku Newtona. Procedura iteracyjna znajdowania punktów stacjonarnych opiera się więc na obliczeniu ze stosunku wartości pierwszej i drugiej pochodnej dla startowego punktu x^k kroku h . Nowa współrzędna x^{k+1} wyznaczana jest z zależności

$$x^{k+1} = x^k + h. \quad (1.50)$$

Przedstawioną wyżej procedurę numeryczną ilustruje rysunek 1.3. Pogrubioną linią ciągłą oznaczono krzywą energii potencjalnej cząsteczki HCl (dla kilku długości wiązania obliczono wartości energii na poziomie teorii DFT/B3LYP/6-311G(d,p) a następnie dopasowano do nich funkcję Morse'a). Za punkt startowy w procedurze iteracyjnej przyjęto $r^1 = 1.5 \text{ \AA}$. Następnie obliczono wartość energii oraz wartości pierwszej i drugiej pochodnej, na podstawie czego wyznaczono równanie paraboli (zaznaczonej linią kropkowaną). Jej minimum w $r^2 = 1.345 \text{ \AA}$ (odpowiadające $h = 0.155 \text{ \AA}$) stanowi lepsze przybliżenie punktu stacjonarnego funkcji Morse'a. Jest to jednocześnie punkt startowy do następnej iteracji, w której buduje się nową parabolę (linia przerywana). Kryteria zatrzymania procedury iteracyjnej zostaną omówione w dalszej części tego rozdziału.

Podobnie dla funkcji wielu zmiennych $f(\mathbf{x})$, gdzie $\mathbf{x} = (x_1, x_2, \dots, x_n)$, rozwinięcie w szereg Taylora drugiego rzędu prowadzi do przybliżenia parabolicznego, tj.

$$f(\mathbf{h}) \approx f(\mathbf{x}^k) + \sum_i g_i h_i + \frac{1}{2} \sum_{ij} h_i h_j H_{ij}, \quad (1.51)$$

gdzie $h_i = x_i - x_i^k$, $g_i = \left(\frac{\partial f}{\partial h_i}\right)_{\mathbf{x}=\mathbf{x}^k}$, $H_{ij} = \left(\frac{\partial^2 f}{\partial h_i \partial h_j}\right)_{\mathbf{x}=\mathbf{x}^k}$. W zapisie macierzowym równanie (1.51) ma postać

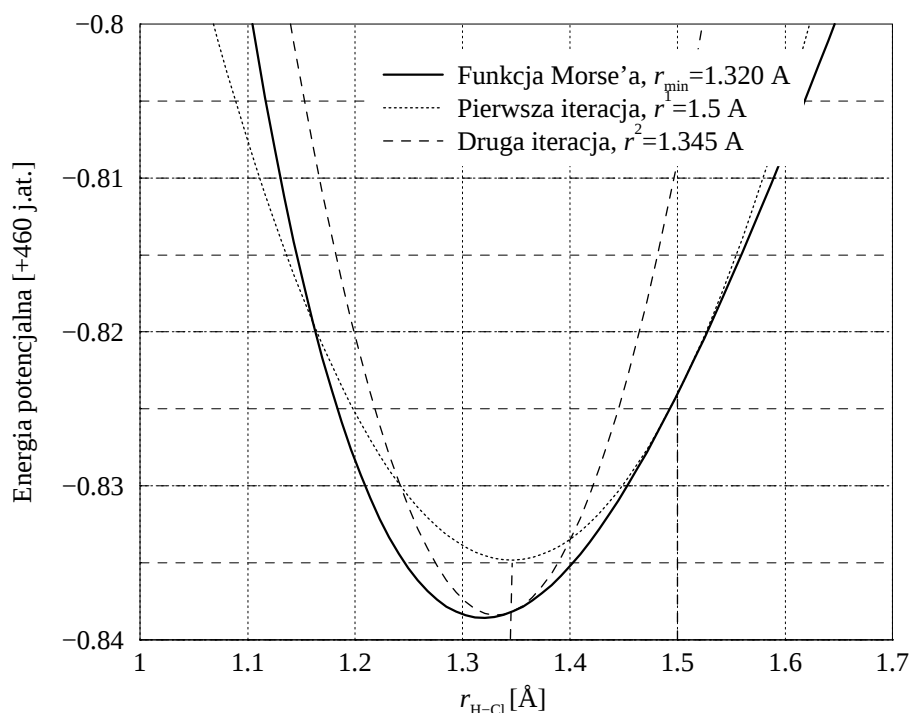
$$f(\mathbf{h}) \approx f(\mathbf{x}^k) + \mathbf{g}\mathbf{h}^T + \frac{1}{2}\mathbf{h}\mathbf{H}\mathbf{h}^T. \quad (1.52)$$

Analogicznie jak w przypadku jednego wymiaru dla punktu stacjonarnego spełniony jest warunek $\frac{\partial f}{\partial h_i} = 0$ dla $i = 1, 2, \dots, n$. Różniczkując równanie (1.52) i korzystając z zależności $\frac{\partial h_i}{\partial h_j} = \delta_{ij}$ otrzymuje się

$$\mathbf{h}^T = -\mathbf{H}^{-1}\mathbf{g}^T. \quad (1.53)$$

Równanie (1.53) jest wielowymiarowym uogólnieniem równania (1.49) (patrz np. [23, 24]). Podobnie także wyznaczane są punkty startowe kolejnych iteracji, tzn.

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{h}. \quad (1.54)$$



Rys. 1.3. Graficzna ilustracja zbieżności metodą Newtona-Raphsona

Procedura iteracyjna zostaje zatrzymana, kiedy dla ustalonych wartości progowych zostaną spełnione podstawowe kryteria zbieżności. Są to:

1. kryterium zmian wartości funkcji. Zmiana wartości funkcji $\Delta f = f(\mathbf{x}^k) - f(\mathbf{x}^{k-1})$ między kolejnymi iteracjami musi być mniejsza od wartości progowej ε_f ;
2. kryterium gradientowe (najważniejsze). Norma¹⁶ gradientu $|\mathbf{g}|$ musi być mniejsza od wartości progowej ε_g ;
3. kryterium długości kroku. Norma kroku $|\mathbf{h}|$ musi być mniejsza od wartości progowej ε_h ;
4. kryterium zbieżności drugiego rzędu. Wielkość $\frac{1}{2}\mathbf{g}\mathbf{H}^{-1}\mathbf{g}^T$, określająca udział członów rzędu drugiego w równaniu (1.52), musi być mniejsza od wartości progowej ε_{s0} .

Kryteria wartości funkcji i gradientowe związane są z testem stacjonarności i wynikają z warunku koniecznego istnienia ekstremum. Często nie są one jednak wystarczające. W przypadku bowiem wystąpienia „płaskiego” minimum zarówno zmiany wartości funkcji jak i gradient mogą przyjmować niewielkie wartości, mimo, że punkt stacjonarny nie został osiągnięty. Z pomocą przychodzi test długości kroku – niewielkie wartości hessianu w opisanej tu sytuacji

¹⁶Normą wektora $\mathbf{a} = (a_1, a_2, \dots, a_n)$ nazywamy jego długość, tj. $|\mathbf{a}| = \sqrt{\mathbf{a}\mathbf{a}^T} = \sqrt{\sum_{i=1}^n a_i^2}$.

powodują, że krok $\mathbf{h}^T = -\mathbf{H}^{-1}\mathbf{g}^T$ jest znaczny, więc procedura iteracyjna nie powinna być zatrzymana. Kryterium drugiego rzędu jest rzadziej stosowane. W praktyce procedurę iteracyjną można zatrzymać wówczas, gdy spełnione są przynajmniej dwa z wymienionych kryteriów. W przypadku cząsteczek za racjonalne wartości progowe przyjmuje się $\varepsilon_f = 10^{-6}$ j.at., $\varepsilon_g = 3 \times 10^{-4}$ j.at. i $\varepsilon_h = 10^{-6}$ j.at. Ciąg punktów generowanych przez algorytm Newtona-Raphsona jest zbieżny z szybkością drugiego rzędu do minimum, o ile w każdym kroku hessian jest ściśle dodatnio określony i spełnia lokalnie warunek Lipschitza (funkcja aproksymująca nie może zbyt szybko rosnać w stosunku do argumentów).