[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 9/45

[12] 发明专利申请公开说明书

[21] 申请号 03145041.5

[43] 公开日 2005 年 1 月 19 日

[11] 公开号 CN 1567200A

[22] 申请日 2003.6.17 [21] 申请号 03145041.5
[71] 申请人 中国科学院长春应用化学研究所
地址 130022 吉林省长春市人民大街 5625 号
[72] 发明人 白汉瀛 董绍俊
权利要求书 2 页 说明书 7 页 附图 1 页

[54] 发明名称 奇异值分解最小二乘法软件

[57] 摘要

在奇异值分解最小二乘(SVDLS)算法原理的基础上，利用 Visual Basic 语言编写而成的数学计算软件，专门应用于解析由圆二色(CD)谱仪测出的圆二色谱的数据。 本软件的优点是良好的人机对话的窗口，良好的操作性，可以避免用户操作上的很多错误，并具有一定的鲁棒性。 点击下拉式菜单开始运算，主对话框中随时提示用户计算的情况，若用户对结果满意，可将之存成磁盘文件，还可将之打开，在主对话框中浏览结果。 将该软件用于解析现场外加电压电解条件下的蛋白质溶液的 CD 谱，得到几种二级结构构象的组分分布。 所得结果与其它方法如 SELCON3 程序相比较，证明是一种快速和有效的方法。

ISSN1008-4274

1. 一种奇异值分解最小二乘法软件软件程序，计算所需的数据文件有两种创建方式，由用户在对话框中进行选择，选择的方式对应两个对话框：一种是在“打开文件”对话框中由用户选定原始数据文件，确定文件组数和每组包含的数据数，由软件在后台将其读入，生成计算用的数据矩阵；另一在“手动输入”对话框中，先在左面的部分由用户确定输入的起始波长和终止波长，而后确定数据的组数，然后在右面部分的输入框中开始一组数据一组数据的输入，输入框的上方还有组数的提示，程序设定手动的输入过程不能反过来进行，输入数据结束后，软件会提示用户将其保存在磁盘上的一个文件中，保存完毕后在主对话框中会自动显示计算的数据矩阵，上面的两个整数分别为数据的组数和每组的数据数。

2. 如权利要求 1 所述的一种奇异值分解最小二乘法软件软件程序，计算数据矩阵创建完成后，利用“文件”菜单中的第二项：“打开数据文件”，从磁盘中打开一个计算数据文件，主对话框中同样会显示其中的数据内容。

3. 如权利要求 1 所述的一种奇异值分解最小二乘法软件软件程序，指定圆二色谱数据中包含的组分数和指定计算的初值，点击“计算”菜单中的“开始”，程序开始计算，程序的线性回归每进行一次，主对话框就会显示计算的结果：偏差 SD 和本次计算与上次计算偏差之间的相对差别，对计算结果，可点击“计算”菜单中的“继续”，

或“完成”。

4. 如权利要求 1 所述的一种奇异值分解最小二乘法软件软件程序，计算结束必须保存结果，点击“文件”菜单中的“保存结果”，将程序内存中的结果保存在一个磁盘文件中，默认结尾为*.end，菜单中的“保存结果”项到现在才可以使用，点击“文件”菜单中的“打开结果文件”，从磁盘上找出结果文件，主界面上就会显示结果：前面为每一个波长下的各个组分的 CD 谱值，后面是每组数据中各个组分的组分数。

奇异值分解最小二乘法软件

技术领域

本发明属于奇异值分解最小二乘法（SVDLS）软件的设计，应用于解析专门由圆二色（CD）谱仪测出的圆二色谱的数据。特别是对于紫外区（185 nm 至 280 nm）区域内各种蛋白质的CD谱。

背景技术

光活性物质及蛋白质圆二色性的起因：当具有光活性的化合物分子的光活性基团处在不对称环境中与偏振光相互作用时就会产生圆二色现象。蛋白质可产生CD谱的空间结构构象会有三种：一级结构，氨基酸残基上的不对称碳原子（除了甘氨酸外），即α－碳上具有的四个不同的取代基；二级结构，肽链骨架中的电子传递和侧链集团的空间螺旋排列；三级结构，对称分子如酪氨酸残基处在蛋白质中的不对称电子环境中。一般的蛋白质的CD光谱分为远紫外区（185~245 nm）和近紫外区（245~320 nm）。远紫外区是肽键的吸收峰范围，许多通常的蛋白质二级结构构象：α－螺旋、平行β－折叠、反平行β－折叠以及β－转角，在此区域都有自己独特的CD谱。近紫外区主要与芳基氨基酸侧链有关，如二硫键。一些氨基酸残基组成的侧链，苯丙氨酸、色氨酸、酪氨酸三种侧链的峰也在此区域内。蛋白质主链的三级结构处在不对称环境中在此区域也会有较强的CD谱，可以作为描

述蛋白质特征的指纹区。

目前用于解析蛋白质圆二色谱，进而分析蛋白质二级结构、三级结构构象的程序软件有很多种。所应用的算法有奇异值分解法，对应的程序名称 SVD；最小二乘法，对应的程序有 MLR、LINCOMB 和 G&F；岭回归法，对应程序名称 CONTIN；顶点限定分析法，对应程序名称 CCA；自洽场分析法，对应程序名称 SELCON3 等等。这些软件只能解析一定区域内的蛋白质的 CD 谱，一般都在 190 nm 至 265 nm 的波长范围内，而且只能解析蛋白质的二级结构构象或三级结构构象。

对于除蛋白质这类生物大分子之外的其它生物或药物小分子，由于它们出现 CD 峰的波长范围常常在可见光区，如 350 nm 至 500 nm，因此到目前为止，还没有合适的 CD 谱分析软件。

发明内容

本发明的目的是设计一种适用于计算解析这些类分子高级构象的软件。

本软件从算法上可以弥补了上述两点不足。一是可以同时在远紫外区分析蛋白质的二级结构构象和在近紫外区分析蛋白质的三级结构构象，二是可以在可见光区分析一些生物小分子的高级结构构象。如胆红素分子，其三级结构构象是两种镜像对称的 M 构象和 P 构象，平时两种构象是外消旋的，即没有 CD 谱；而当受外界不对称环境诱导时，一种构象减少，另一种增加，即会产生较大的 CD 峰。

在 n 个波长下对 m 个样品进行测定，得到 $n \times m$ 个光的吸收值。如果样品中含有 k 个组分，且其光吸收具有加合性并服从

Lamber-Beer 定律，则在 1 cm 光程长度下每一吸光度值可表示为，

$a_{i,j}=\sum\varepsilon_{i,l}c_{j,l}(i=1,\cdots,n;\ j=1,\cdots,m;\ l=1,\cdots,k)$

其中$\varepsilon_{i,l}$为 l 组分在第 i 处波长的摩尔吸收系数，$c_{j,l}$ 为 j 样品中第 l 个组分的浓度。将吸光度组成吸收矩阵并用黑体大写字母表示，$\mathbf{A}_{n\times m}$。根据奇异值分解的数学原理，A 可分解为三个矩阵相乘，

$\mathbf{A}=\mathbf{USV}$

其中 U 为 $n\times n$ 正交矩阵，V 为 $m\times m$ 正交矩阵，S 为 $n\times m$ 对角矩阵，$\mathbf{S}^2$ 为矩阵 A 的 m 个特征值按降序排列。按照因子分析原理，由 k 个组分组成的样品的吸收矩阵 A 的前 k 个特征值（$k<m$）称为主因子，可用于描述矩阵 A 的主要特征。而 k 之后的特征值是由于测量等误差造成的。用 U 的前 k 列，S 和 V 的前 k 行组成新的矩阵，$\overline{\mathbf{E}}_{n\times k}$ 和 $\overline{\mathbf{C}}_{k\times m}$，两者之积称为抽象光谱矩阵 $\overline{\mathbf{A}}_{n\times m}$，

$$\overline{\mathbf{A}}_{n\times m}=\overline{\mathbf{E}}_{n\times k}\overline{\mathbf{C}}_{k\times m} \qquad (1)$$

其中 $\overline{\mathbf{E}}_{n\times k}=\mathbf{U}_{n\times k}\mathbf{S}_{k\times k}$ 称为 k 个光谱的抽象吸收系数矩阵，$\overline{\mathbf{C}}_{k\times m}$ 称为 m 样品中 k 个组分的抽象浓度矩阵。如何将 $\overline{\mathbf{E}}$ 和 $\overline{\mathbf{C}}$ 转化为真实的吸收系数矩阵和浓度矩阵是该方法的关键。以实验获得的吸收矩阵为依据，以抽象吸收系数矩阵为基础，以 $c_{k,m}\geq 0$ 为条件，以标准偏差达到稳定为判据，通过最小二乘法将抽象吸收系数矩阵和浓度矩阵转化为具有物理意义的矩阵。根据最小二乘法原理，

$$\overline{\mathbf{E}}=\mathbf{A}\overline{\mathbf{C}}^{\mathrm{T}}(\overline{\mathbf{C}}\overline{\mathbf{C}}^{\mathrm{T}})^{-1}$$

$$\overline{\mathbf{C}}=(\overline{\mathbf{E}}^{\mathrm{T}}\overline{\mathbf{E}})^{-1}\overline{\mathbf{E}}^{\mathrm{T}}\mathbf{A}$$

由式（1）获得的 $\overline{\mathbf{A}}$ 与 A 的标准偏差为，

$$SD=\sum\sum[(A-\bar{A})/(n\times m-k)]^{1/2}$$

当 SD 趋于恒定时，所获得的 $\bar{E}=E$，$\bar{C}=C$。

利用 Microsoft 公司的 Visual Basic 语言编译器可以创作各种可视化窗口的强大功能，编写可视化操作界面和计算程序内核，编制成为界面友好、操作简单的数据处理软件。

计算程序的流程图如图 1 所示。

平行四边形框中的“数据输入”和“数据输出”由软件的对话窗口完成。矩形框中的各个程序部分为本软件完成运算必需的程序内核，圆角矩形框的程序部分为可选择性的程序内核。这些程序内核为本软件所独有，尤其是“奇异值分解运算”和“最小二乘法处理”部分。

本软件最大的优点是良好的人机对话的窗口，良好的操作性，能够防止用户的很多误操作。

数据文件分为原始的数据文件——即从 CD 谱仪测得的一组一组的数据，和用于计算的数据文件——将若干组原始数据组合成为计算所需的数据矩阵。软件主界面为下拉式菜单，包含数据创建、数据的 I/O、组分设定、初值设定、计算、版权声明和帮助文件这些所有的指令。软件体中心为主对话框，用来显示输入的数据、计算的过程和最后的结果。程序的运行顺序为步进式，必须先运行前面的数据创建或数据的输入步骤，才能运行组分设定和初值设定，进而才能运行计算的过程，最后才能保存和显示计算的结果，这样就增强了程序运行的鲁棒性（robust）。

计算数据文件的创建方式有两种，由用户在对话框中进行选择，选择的方式对应两种对话框：一种是在“打开文件”对话框中由用户选定原始数据文件，确定文件组数和每组包含的数据数，由软件在后台将其读入，生成计算用的数据矩阵；另一在“手动输入”对话框中，先在左面的部分由用户确定输入的起始波长和终止波长，而后确定数据的组数，然后在右面部分的输入框中开始一组数据一组数据的输入，输入框的上方还有组数的提示。程序设定手动的输入过程不能反过来进行，这样可以避免用户在输入过程出现的一些不经意的错误（多输或少输）。输入数据结束后，软件会提示用户将其保存在磁盘上的一个文件中，保存完毕后在主对话框中会自动显示计算的数据矩阵，上面的两个整数分别为数据的组数和每组的数据数。

计算数据矩阵创建完成后，为方便用户打开一个已经创建的计算数据文件，利用“文件”菜单中的第二项：“打开数据文件”，从磁盘中打开一个计算数据文件，主对话框中同样会显示其中的数据内容。

随后的过程是指定 CD 谱数据中的组分数，和指定计算的初值。

点击“计算”菜单中的“开始”，程序开始计算过程，线性回归每进行一次，主对话框就会显示计算的结果：偏差 SD 和本次计算与上次计算偏差之间的相对差别。如果用户对计算结果不满意，可点击“计算”菜单中的“继续”，令计算过程继续进行，若满意，则点击“计算”菜单中的“完成”，令计算结束。

计算结束必须保存结果，点击“文件”菜单中的“保存结果”，将程序内存中的结果保存在一个磁盘文件中，默认结尾为*.end。菜

单中的“保存结果”项到现在才可以使用，这也是防止操作顺序出错。若想查看结果，点击“文件”菜单中的“打开结果文件”，从磁盘上找出结果文件，主界面上就会显示结果：前面为每一个波长下的各个组分的 CD 谱值，后面是每组数据中各个组分的组分数。

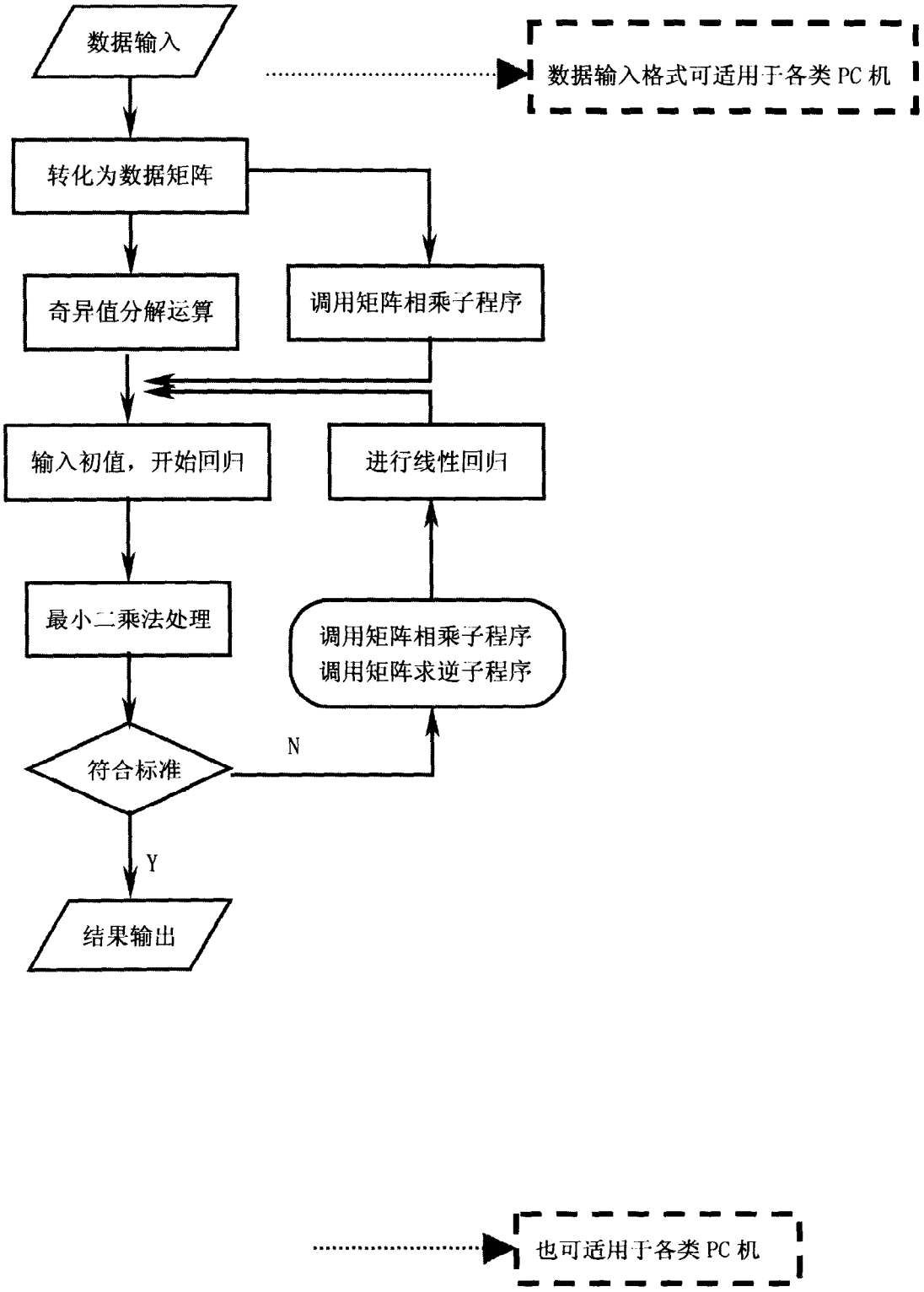
具体实施方式

用编制好的 SVDLS 软件分析本人所作的外加电场下的牛血清白蛋白与胆红素小分子相互作用的 CD 现场光谱电化学实验数据，得到图 2 的结果，所得的牛血清白蛋白的二级结构构象分析结果，再与专门分析蛋白质二级结构的程序 SELCON3 作比较，结果比较见表 1。可以发现，两者的差异很小。说明本软件分析蛋白质的二级结构的结果非常良好。

再用本软件分析胆红素分子在牛血清白蛋白的诱导作用下的 CD 现场光谱电化学实验数据，可得到图 3 的结果，可以清晰地看到胆红素分子两种三级结构构象组分随电场变化的曲线。

表 1　SVDLS 与 SELCON3 两种程序分析同一 CD 数据的结果对比

外加电位/V 组分		0.00	0.10	0.17	0.24	0.31	0.38	0.45	0.52	0.60
α-螺旋	SVDLS	0.46	0.53	0.60	0.62	0.69	0.75	0.77	0.71	0.65
	SELCON	0.42	0.49	0.50	0.58	0.64	0.72	0.76	0.72	0.72
平行β-折叠	SVDLS	0.17	0.16	0.14	0.14	0.09	0.06	0.05	0.05	0.05
	SELCON	0.18	0.14	0.14	0.13	0.09	0.07	0.05	0.05	0.04
反平行β-折叠	SVDLS	0.12	0.11	0.10	0.10	0.09	0.07	0.07	0.06	0.08
	SELCON	0.12	0.11	0.11	0.10	0.09	0.08	0.07	0.07	0.07
β-转角	SVDLS	0.07	0.06	0.07	0.04	0.02	0.04	0.01	0.01	0.02
	SELCON	0.07	0.07	0.06	0.05	0.03	0.03	0.01	0.02	0.01
无规则卷曲	SVDLS	0.23	0.14	0.20	0.15	0.16	0.11	0.10	0.15	0.16
	SELCON	0.21	0.19	0.19	0.105	0.15	0.10	0.10	0.13	0.15


数据输入
数据输入格式可适用于各类 PC 机
转化为数据矩阵
奇异值分解运算
调用矩阵相乘子程序
输入初值，开始回归
进行线性回归
最小二乘法处理
调用矩阵相乘子程序
调用矩阵求逆子程序
符合标准
N
Y
结果输出
也可适用于各类 PC 机