

分类号 TP391 密级 公开

UDC 学号 20120613014

青海师范大学

硕士学位论文

基于 MATLAB 藏药 X 射线衍射图谱的数据处理

研究生姓名 刘远

导师姓名（职称） 段新文， 教授

申请学位类别 理学 申请学位名称 硕士

学科专业名称 计算机应用技术 研究方向名 EDA 与计算机仿真技术

论文提交日期 2015 年 4 月 论文答辩日期 2015 年 5 月

学位授予单位 青海师范大学 学位授予日期 2015 年 6 月

答辩委员会主席 谢孟荣

评阅人 宋传鸣， 郭敏

基于 MATLAB 藏药 X 射线衍射图谱的数据处理

中文摘要

X 射线衍射指纹图谱法现已成为现代中药分析领域的重要方法之一，而对于藏药而言，仍存在诸多问题，其中之一便是藏药材 X 射线衍射图谱的数据处理。大部分的藏药材，尤其是植物类，由于含有淀粉、蔗糖、蛋白质等多属低级晶系的大分子有机物质，结晶度很低，毛刺较多，其 X 射线衍射图谱整体上呈模糊弥散型宽峰，特征衍射峰较少又叠加其上，很难进行图谱解析，也就限制了其在藏药材真伪鉴定和质量控制上的应用。本文正是以解决该问题为目的展开一些研究工作的，具体内容包括四个方面：

一是，基于现有的像 MDI jade5.0 和 X' Pert Highscore Plus 等专业 X 射线衍射图谱分析软件介绍了格式转化、图谱平滑、背景及 $K\alpha$ 扣除、图谱寻峰等传统数据预处理功能模块在 Jade5.0 中的实施过程及相关算法，还重点对 Savitzky-Golay 平滑滤波器、小波模极大值去噪算法和小波多区间阈值去噪算法在植物类药材 X 射线衍射图谱去噪效果方面进行了对比研究。

二是，提出获取藏药材尤其是植物类藏药粉末衍射文件的方法，以解决植物类藏药材 X 射线衍射图谱难以解析。该方法包括获取 X 射线衍射图谱文件、获取二阶导数图谱文件、寻峰处理、制作粉末衍射文件等四个主要步骤，分别采用了 Savitzky-Golay 平滑滤波器和对称零面积卷积寻峰算法。在该部分，我们对三点公式法、五点公式法、三次样条法和三次 B 样条法等离散数据二阶导数计算方法进行了原理上的介绍。基于 MATLAB 编程语言，我们编写了一套程序，能批量获取藏药材 X 射线粉末衍射文件的相关信息，方便快捷。

三是，提出了植物类藏药材 X 射线衍射图谱分峰系统。该系统包括格式转化、图谱平滑、背景扣除、初步寻峰和分峰拟合等五个子系统模块，涉及到小波模极大值去噪、统计敏感的非线性迭代削峰、对称零面积卷积寻峰、最小二乘拟合等算法。文中给出了各算法实施步骤。基于 MATLAB 语言和上述算法，编写了 X 射线衍射图谱分峰软件，其具备交互式图形界面，能有效地对图谱进行分峰处理，获得峰位、峰强、半高宽、峰型因子、偏态因子等丰富的衍射峰信息，为进一步的药材分析如结晶度计算、相似度计算提供了参考依据。

四是，基于向量夹角法、相关系数法、相对熵理论，提出了 X 射线衍射图谱相似度计算模式，包括原始数据采集、标准化处理、特征信息提取和相似度计算等四个步骤，其中特征信息提取部分有三种方式可以获取。一是，利用离散数学中的二阶导数理论对原始图谱进行信息挖掘，获得足够的指纹特征；二是，引入

Voigt 衍射峰函数模型和全谱拟合技术,对原始图谱进行分峰拟合,获得精确独立的衍射峰信息;三是直接应用原始图谱,结合相对熵算法,获得图谱间相似度。

以上便是本文的主要研究内容和成果,总的来说,提出了获取粉末衍射文件的流程、藏药材 X 射线衍射图谱分峰系统和图谱相似度计算模式等三种方法用以解决植物类藏药材 X 射线衍射图谱很难解析的问题。相信上述研究对实现藏药材的高效准确快速鉴定、质量控制和藏药 X 射线衍射指纹图谱数字化必将起到积极的作用。

关键词: 藏药, X 射线衍射, MATLAB, 数字分峰, 图谱相似度, 二阶导数图谱

Data Processing for X-ray Diffraction Pattern of Tibetan Medicine Based on MATLAB

Abstract

X-ray diffraction(XRD) having been an important method involved in modern traditional Chinese medicine analysis, some problems yet exist for Tibetan medicine, one of which lies in the data processing of XRD patterns. The fingerprints of most of Tibetan medicine, especially the plant-based, exhibit a vague disperse broad peak with massive ‘burr’ and fairly few characteristic diffraction peaks overlapping, owing to the medicine containing some macro-molecular organic substances such as starch, saccharose, protein, and so on, belonging to elementary syngony, which leads to fairly difficult patterns analysis and restrain its application on identification and quality control of Tibetan medicine. Some researches conducted in this paper are just aimed at resolving this problem, and the detailed contents include four main parts as follows. For the first part, based on professional XRD pattern analysis software such as MDI Jade5.0 and X’Pert Highscore Plus, the implementation in Jade5.0 and relevant algorithms of traditional data pre-processing modules such as format transformation, pattern smoothing, background & $K\alpha$ removal and peak searching, were introduced. A comparison study on denoising effects of XRD patterns was conducted between Savitzky-Golay filter, wavelet Modulus Maximum denoising algorithm and multi-interval threshold denoising algorithm.

For the second part, a procedure utilized for obtaining powder X-ray diffraction files was put forward to resolve the problem of difficult pattern analyzing of plant-based Tibetan medicine, which encompasses four main steps, that is, gaining XRD pattern files, gaining second derivative pattern files, peak searching and making powder X-ray diffraction files, in which Savitzky-Golay filter and symmetric zero-area convolution peak searching algorithm were used. In this part, we made a brief introduction in principle for second derivative patterns calculating of discrete data such as three-points formula method, five-points formula method, cubic spline method and cubic B-spline method. Based on MATLAB, we developed a set of programs, which can obtain relevant information of Tibetan medicine powder X-ray diffraction files, conveniently and fast.

For the third part, the X-ray Diffraction Whole Pattern Peak-resolution Software of Plant Tibetan Medicine was designed, which includes five subsystem modules of format transformation, pattern smoothing, background removal, initial peak searching and peak-resolution fitting, covering wavelet Modulus Maximum denoising algorithm, sensitive nonlinear iterative clipping algorithm and least square fitting. Based MATLAB and the above algorithms, the software for making XRD pattern peak-resolution was developed, which possesses interactive graphic interface and is

quite easy for user to do peak-resolution work and obtain abundant diffraction peaks information consisting of peak locations, peak intensity, full width at half maximum(FWHM), peak-shape factor and skewness factor, providing reference for further medicine analysis such as crystallinity calculation and similarity calculation. For the forth part, based on vectorial angle method, correlation coefficient method and relative entropy theory, a similarity calculation procedure was given for XRD patterns, including four steps of original data collection, normalization, feature information extraction and similarity calculation.

The mentioned above are just the main research contents and consequences. In general, three methods, that is, powder X-ray diffraction files gaining procedure, X-ray Diffraction Whole Pattern Peak-resolution Software of Plant Tibetan Medicine and pattern similarity calculation mode were studied and designed for the XRD pattern analyzing of Tibetan medicine, especially the plant-based. It is believed that the research conducted in this paper will play an active role in performing efficient fast accurate identification and quality control of Tibetan medicine and the XRD patterns' digitalization of Tibetan medicine.

Keywords: X-ray diffraction; MATLAB; digital peak-resolution; pattern similarity; second derivative pattern

目 录

第一章 绪论.....	1
1.1 背景介绍.....	1
1.2 研究内容.....	2
1.2.1 二阶导数图谱.....	2
1.2.2 数字分峰技术.....	2
1.2.3 相似度计算.....	2
1.3 研究手段.....	3
1.3.1 粉末 X 射线衍射分析.....	3
1.3.2 MATLAB.....	3
第二章 XRD 数据处理及相关算法理论.....	4
2.1 格式转化.....	4
2.2 平滑处理.....	7
2.3 背景及 $K\alpha$ 扣除.....	9
2.4 图谱寻峰.....	11
第三章 X 射线衍射二阶导数指纹图谱.....	13
3.1 二阶导数指纹图谱的数学基础.....	13
3.1.1 三点公式法.....	15
3.1.2 五点公式法.....	18
3.1.3 三次样条法.....	20
3.1.4 三次 B 样条法.....	22
3.2 植物类药材粉末衍射文件.....	23
3.2.1 获取 PDF 文件的方法.....	23
3.2.2 应用实例.....	24
第四章 藏药 XRD 全谱分峰系统.....	27
4.1 章节引言.....	27
4.2 XRD 图谱数字分峰理论体系.....	29
4.2.1 小波去噪.....	29
4.2.2 背景扣除.....	29
4.2.3 对称零面积卷积寻峰.....	30
4.2.4 全谱拟合技术与数字分峰.....	31
4.3 程序开发与应用实例.....	31
4.3.1 程序开发.....	31

4.3.2 应用实例.....	33
第五章 藏药 XRD 二阶导数指纹图相似度计算.....	36
5.1 相似度理论及算法.....	36
5.1.1 夹角余弦法.....	36
5.1.2 相关系数法.....	37
5.1.3 相对熵方法.....	37
5.2 相似度计算模式及程序设计.....	39
5.3 相似度应用实例.....	39
5.3.1 样品与实验.....	40
5.3.2 贝壳类矿石藏药.....	40
5.3.3 不同产地大黄图谱相似度分析.....	43
第六章 总结与展望.....	46
6.1 成果与收获.....	46
6.2 展望.....	46
参考文献.....	47
致谢.....	49
在校期间的研究成果及发表的学术论文清单.....	50

第一章 绪论

1.1 背景介绍

藏药是我国传统药学的一个组成部分，目前有药用记录的藏药达 2294 种，其中常用的 300 多种，植物类 200 余种，占 70%，动物类 40 余种，占 12%，矿物类 40 余种，占 14%，和中药交叉使用的药材就有 274 种。

由于历史的局限性，传统的藏药生产、加工炮制的方法和技术较为混乱，剂型简单，使用方法因人而异。为振兴藏医药，国家投入大量人力物力对藏医药进行了系统整理和发掘研究，借鉴和应用现代科学技术和研究手段是推动藏药走向现代化的关键。现有的研究手段包括紫外光谱、红外光谱、高效液相色谱法和 X 射线衍射（XRD）等技术方法。这里除了 XRD，其他几种分析方法已广泛用于藏药材成分测定和质量控制，如杨红霞等人^[1]应用红外光谱技术和二级导数光谱技术对藏药川西獐牙菜不同提取物的药用成分进行分析研究，以达到对后期药理活性成分的宏观上的控制；刘震东等人^[2]采用薄层色谱法和高效液相色谱法测定诃子等药材中没食子酸的含量，并以此建立鉴别方法；吴红彦等人^[3]也利用这两种技术方法作为石榴健胃片中红花、肉桂的鉴定方法，以控制石榴健胃片的质量。

在传统中药分析领域，人们应用 X 射线衍射手段尤其是粉末 X 射线衍射分析法做了很多的工作，为后续药学研究者们提供了丰富的经验。如吕扬等人^[4]以茜草等多种中药样品为例，对他们的 XRD 图谱进行观察分析，发现借助从采集得到的图谱中挖掘出来的特征衍射峰数据和几何拓扑规律，建立具备“指纹功能”的特性信息，能够实现中药材的真伪鉴别和质量评价。而周国俊等人^[5]应用粉末 X 射线衍射技术研究了采自不同地区的中药蛇床子，发现不同产地的样品，其图谱几何拓扑呈现出 4 种相似的衍射图形，表明该技术可以用来进行中药鉴定和归类。王树春等人^[6]通过观察分析药品的 X 衍射 Fourier 谱，发现其中的差异和规律，进而用以鉴别真伪熊胆和不同种类的熊胆。郑笑为等人^[7]应用粉末 X 射线衍射测试手段对多种珍珠进行了主成份分析，从而鉴定珍珠种类，也对其质量进行了评价。上述研究表明了 X 射线衍射技术在药材鉴定、归类和评价方面具备很好的应用前景。

和中药相比，XRD 法在藏药分析领域的应用就很少了，已有的研究也只是对藏药材的定性鉴定，即通过识别特征衍射峰来鉴别藏药。如李岑等人^[8]采用 X 射线衍射对不同来源珠西的结构组成进行测定，以期揭示药材药理的物理基础；赵旭东等人^[9]采用粉末 X 射线衍射法对不同种大黄样品进行分析，获得唐古特大黄

特征衍射峰值，建立标准的 X 射线衍射图谱；全正香等人^[10]对藏药南寒水石进行了物相分析。而且，无论对于藏药材还是中药材，还存在另一个问题，即：传统的 XRD 数据分析模式，对于矿物类和化石类药材较为有效，而对于植物类药材已不再适用。这是因为植物类药材中含有大量有机成分，如淀粉、蔗糖、纤维等，它们多属低级晶系，因而结晶度较低，图谱毛刺很多，整体呈模糊弥散型宽峰。

本论文中所提出的二阶导数图谱方法和数字化 XRD 全谱分峰软件能有效解决植物类藏药 X 射线衍射图谱难以解析的问题；联合提出夹角余弦法、相关系数法和相对熵理论用于 XRD 图谱相似度分析；这两方面的工作将会丰富 XRD 图谱在藏药质量鉴别和控制方面的应用。

1.2 研究内容

1.2.1 二阶导数图谱

对于植物类藏药材而言，XRD 图谱毛刺太多，整体呈弥散型峰包，传统的 XRD 分析方法极难对其进行表征，也就不能进一步用以药材真伪鉴定和质量控制。为此，我们研究二阶导数图谱技术来对此类图谱进行解析。二阶导数图谱计算属于数值分析领域。这里将要对三点法、五点法、三次样条法、三次 B 样条法等六种二阶导数图谱计算方法进行原理介绍和 MATLAB 程序编程。

1.2.2 数字分峰技术

除了二阶导数图谱法外，数字分峰技术也能有效解决植物类药材 XRD 图谱解析问题。所谓数字分峰技术就是多峰重叠的衍射峰包分解成相应的独立衍射峰，获得精确的衍射角度、衍射强度和半高宽等特征数据。这里我们尝试建立一个植物类 XRD 图谱分峰系统，涉及 XRD 图谱文件格式转化、图谱平滑、背景去除、分峰拟合等多个子技术部分。

1.2.3 相似度计算

相似度计算一般采用两种方法进行，即相关系数法和向量夹角余弦法。利用相关系数可确定两种属性之间的关系，它强调了数值涨落的比较。而向量夹角法强调的是图谱间整体上的相似性。特别地，相对熵理论也可以表征图谱间的相似性。相对熵又称为 KL 散度，它是对两个概率分布差异性的非对称性的度量。文中将分别以衍射角度和衍射强度的角度研究如何应用相对熵对 XRD 图谱进行相

似度表征。

1.3 研究手段

1.3.1 粉末 X 射线衍射分析

和单晶 X 射线衍射不同,粉末 X 射线衍射实验需要将样品粉碎成很细的粉末才可以使用。应用粉末 X 射线衍射可以做很多工作,如对固态物质进行物相分析,研究内部结构与材料宏观性能间的关系;测定材料的晶体结构以及与之相关的参数;测定材料的微观应力;也能进行织构分析;对于结晶程度不高的物质比如植物类藏药,它也能在一定程度上对其结构进行表征。粉末 X 射线衍射实验采集到的指纹图谱,是对材料组份、晶体机构、元素类型等的反应,具有很高指纹特性。本文使用的衍射仪器是荷兰飞利浦公司生产的 Y-4Q 型全自动粉末 X 射线衍射仪,用来做实验的药品是采自青海、甘肃、四川、西藏等地的藏药材。在进行 XRD 实验前,我们首先需要对药材进行 70℃烘箱干燥处理,然后用粉碎机进行粉碎,过 120 目筛子,最后压片实验。本文 XRD 实验的测试参数是:阳极材料, Cu-K α ; 滤波材料, Ni; X 射线波长, 0.154178nm; 管电流, 20mA; 管电压, 30kV; 扫描速度, 0.3°/s; 扫描范围, 20~90°。

1.3.2 MATLAB

MATLAB 是一款功能非常强大的高级编程语言,由公司 MathWorks 开发。在工程计算、图像处理、信号仿真以及科学计算领域能经常看到它的影子,尤其是在科学计算领域,它包含众多专业性很强的工具箱可供使用。MATLAB 是一门高级语言,也就是用户不用费心考虑计算机的底层实现,只需关注问题的解决方案,这样大大提高了工作效率。而且它具备很好的数据可视化功能,用户能够直观的面对计算结果,方便调试,也增加了编程带来的乐趣。MATLAB 内置了大量的函数,用户有时只需要一个调用就能实现自己想要的结果。文中,我们经常采用 MATLAB 脚本来编写 XRD 相关数据处理程序,也在 GUI 平台上设计具备交互式功能的图谱分峰系统。

第二章 XRD 数据处理及相关算法理论

由于一些藏药材 XRD 图谱毛刺较多,背景很高,整体呈弥散型宽峰,衍射峰隐没其中,因而传统的 XRD 分析软件很难对其进行解析,需要建立新的数据处理模式,开发独立于现有 XRD 专业分析软件之外的程序及相关算法。本文要建立三种数据处理模式,分别是藏药材 PDF 制作系统,用于获取藏药材衍射数据和一些基本信息,为建立藏药 XRD 图谱专家系统提供数据基础;藏药材 XRD 图谱数字分峰系统,用于对植物类弥散型 X 射线衍射图谱进行分峰处理,以供药材鉴定和相似度计算之用;藏药材相似度计算系统,用于计算相近药材 XRD 图谱间的相似度,以便鉴定和归类。他们都是基于传统的 XRD 分析模式建立起来的。

现有的用于 XRD 图谱分析的软件有 JADE5.0, X'Pert Highscore Plus, PowderX, Search Match 等,其中 MDI Jade 是由美国材料数据公司开发的,在国内应用较为广泛,目前普遍使用的版本是 Jade5.0,功能非常强大,可以进行衍射峰的指标化、晶格参数的校正、晶格参数的计算、衍射峰面积和质心的计算等等,而且其出图功能相当强大,用户可以进行更加随意的编辑。X'Pert Highscore Plus 是帕纳科公司开发的,它也非常实用、功能强大的 XRD 常规分析软件,尤其是在自动检索方面较为突出。但是无论何种分析软件,他们在图谱的预处理模块方面都是相似的,都包含数据平滑、背景及 $K\alpha$ 扣除、寻峰等。下面结合 MDI Jade5.0,对图谱预处理的各个模块进行介绍,并引入一些新的处理算法,为后面三章开展其他模式开发奠定基础。

2.1 格式转化

使用的 X 射线衍射仪不同,最后产生的文件格式也是不同的。比如本文使用的荷兰飞利浦公司生产的 Y-4Q 型全自动粉末 X 射线衍射仪,其采集得到的是以 mdi 作为扩展名的文件,其内部数据结构见图 3-1。可以看到该文件包含文件头信息(A,B 标示)和主体数据部分(C 标示)。文件头部分,第一行 A 是仪器测试条件,第二行 B 代表 XRD 图谱采集参数:衍射开始角 3.0° ,衍射角步长 0.03° , $K\alpha$ 辐射 Cu 靶,X 射线平均波长 1.54178\AA ,衍射终止角 90° ,衍射数据个数 2901 个。剩下的 C 部分就是 X 射线衍射数据,有效个数为 2901 个,每行 8 个数据,中间用 5 个空格隔开。

现在需要将此类 mdi 格式文件转化为 txt 格式的文件,文件里只包含两列数据,一列代表衍射角度,一列代表衍射强度。这样做便于用各个专业 XRD 分析软件分析使用,也方便自编程序读取数据,进行一些批量处理。像 JADE5.0 和

Highscore 等专业的 XRD 分析软件，都能进行格式转化，但都不能进行批量的格式转化操作。这里以 JADE5.0 为例介绍格式转化的过程。首先，打开软件读取 mdi 格式文件；然后点击 File→Save→Primary Pattern as *.txt，继而会弹出保存文件窗口；最后再保存相应文件即可。打开转化后的 txt 文件发现文件仍含有代表测试条件的文件头信息。下面我们应用 MATLAB 语言编写格式转化程序，如下：

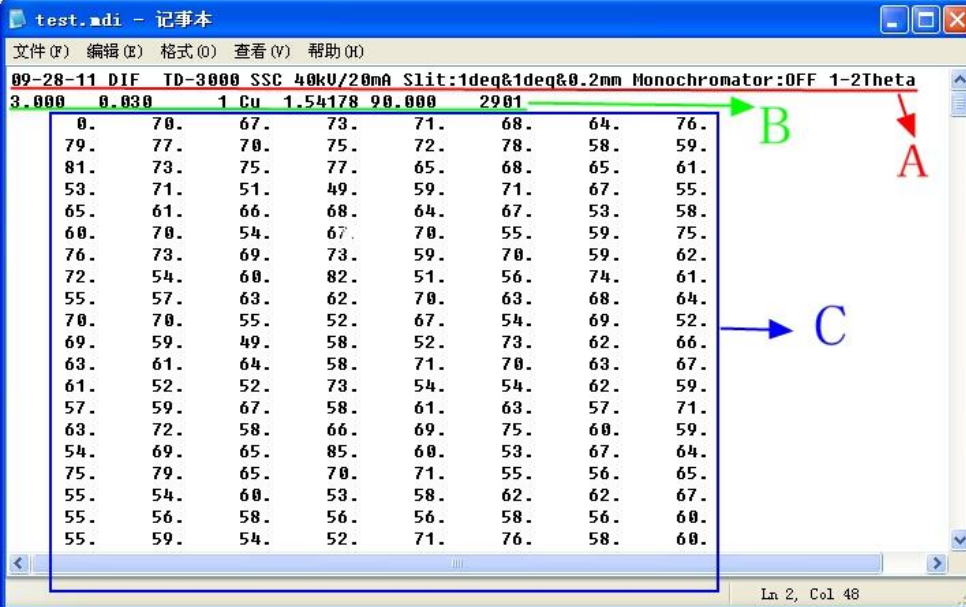
```
clear all; close all; clc;
%批量格式转化
files = dir('D:\我的文档\MATLAB\批量获取二阶导数指纹图谱并画图\test\原始MDI 图谱\*.mdi');
N = numel(files);
%设定各类型文件存储路径
dir_txt = 'D:\我的文档\MATLAB\批量获取二阶导数指纹图谱并画图\test\TXT 格式图谱';
for i=1:N
    %格式转化
    mdi_name = files(i).name;
    [Int_ori,Theta_ori] = FormatTrans(mdi_name);
    filename = strrep(files(i).name, '.mdi', 'new.txt');
    pattern_txt = [Theta_ori Int_ori];
    dlmwrite([dir_txt filename],pattern_txt,'delimiter','\t','newline','pc');
end
function [Int,Theta] = FormatTrans(fullname)
%%=====
% 该程序用于将 MDI 格式 XRD 图谱文件转化为 TXT 格式文件
% 输入：待进行格式转化的 MDI 文件的全路径名 fullname
% 输出：保存 TXT 格式文件
%%=====
%打开待转化 MDI 文件
fid=fopen(fullname,'r');
%读取文件前两行的文字信息及后面的 XRD 强度数据
C=textscan(fid,'%f %f %d %s %f %f %d',1,'HeaderLines',1);
%HeaderLines：意为跳过开头几行
%1：意为跳过开头 1 行
```

```

%C: 读取的信息以元胞方式存在
fstart=C{1}; fstep=C{2}; fstop=C{6};
Int = fscanf(fid,'%d. ');
Theta = fstart:fstep:fstop;
Theta = Theta';
Int = Int(1:length(Theta));
%强度归一化
Int = Int./(max(Int));
%关闭文件
fclose(fid);
%保存TXT 格式文件
Theta = Theta(2:end); Int = Int(2:end);
end

```

应用以上程序可以将 mdi 格式文件批量转化为 txt 格式，转化后的文件见图 2-2，可以看到文件里只有两列数据，L 列为从 3~90°范围内，间隔 0.03°的衍射角数据；R 列为与衍射角相对应的衍射强度数据。



09-28-11 DIF TD-3000 SSC 40kV/20mA Slit:1deg&1deg&0.2mm Monochromator:OFF 1-2Theta

3.000	0.030	1	Cu	1.54178	90.000	2901	
0.	70.	67.	73.	71.	68.	64.	76.
79.	77.	70.	75.	72.	78.	58.	59.
81.	73.	75.	77.	65.	68.	65.	61.
53.	71.	51.	49.	59.	71.	67.	55.
65.	61.	66.	68.	64.	67.	53.	58.
60.	70.	54.	67.	70.	55.	59.	75.
76.	73.	69.	73.	59.	70.	59.	62.
72.	54.	60.	82.	51.	56.	74.	61.
55.	57.	63.	62.	70.	63.	68.	64.
70.	70.	55.	52.	67.	54.	69.	52.
69.	59.	49.	58.	52.	73.	62.	66.
63.	61.	64.	58.	71.	70.	63.	67.
61.	52.	52.	73.	54.	54.	62.	59.
57.	59.	67.	58.	61.	63.	57.	71.
63.	72.	58.	66.	69.	75.	60.	59.
54.	69.	65.	85.	60.	53.	67.	64.
75.	79.	65.	70.	71.	55.	56.	65.
55.	54.	60.	53.	58.	62.	62.	67.
55.	56.	58.	56.	56.	58.	56.	60.
55.	59.	54.	52.	71.	76.	58.	60.

图 2-1 mdi 格式文件



3.0	0
3.03	70
3.06	67
3.09	73
3.12	71
3.15	68
3.18	64
3.21	76
3.24	79
3.27	77
3.3	70
3.33	75
3.36	72
3.39	78
3.42	58
3.45	59
3.48	81
3.51	73
3.54	75

图 2-2 转化后的 txt 格式文件

2.2 平滑处理

前面已经叙述，一些藏药材因为结晶度很低，XRD 图谱毛刺很多，图谱显得很“平滑”，这阻碍了进一步的分峰和寻峰。因此需要应用一定平滑滤波算法对其进行处理。Jade5.0 应用 Savitzky-Golay 最小二乘滤波器在角度域内对图谱进行平滑处理。在 Jade5.0 软件中用户可以通过右击主工具栏上的滤波按钮调出图谱平滑参数设置窗口，见图 2-3。当改变该窗口上的参数时，Jade 会立刻对图谱按相应参数设定进行平滑，然后将平滑后的图谱叠加显示在缩放窗口中，以便对比。用户通过左右拖拽滑块在 5~99 个点间改变滤波器窗口宽度，但要注意选择的点数越大，虽然图谱可能比较平滑但是峰型失真也会很严重。这里提供两类滤波器：抛物线滤波和四次滤波，后者在保留衍射峰肩方面要优于前者。当选择较长滤波器来平滑宽峰和背景时，如果选择 Smooth and Preserve Peaks，Jade 会保留图谱中的尖峰。如果选择 Smooth Background，Jade 就只平滑图谱背景区域。

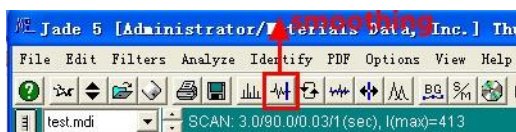


图 2-3 Jade5.0 主工具栏

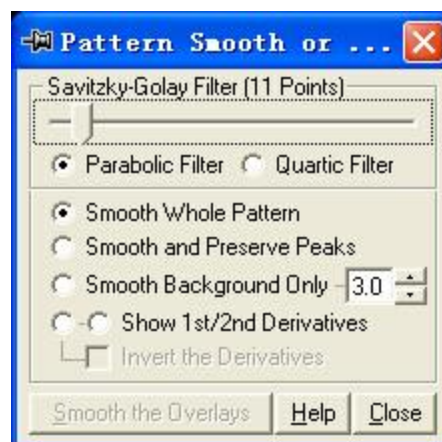


图 2-4 图谱平滑参数设置窗口

以上就是应用 Jade5.0 进行平滑处理的简要介绍, 其中涉及到的核心算法是 Savitzky-Golay 滤波器。它是一种数字滤波器, 用于对一组数据点进行平滑处理, 即在不造成大的信号失真情况下提高信噪比值。Savitzky-Golay 平滑滤波器实质上是一个卷积过程: 基于最小二乘法, 应用一个低阶多项式对连续的相邻数据子集进行拟合。当数据点等距, 我们就可以获得最小二乘方程组的一个分析解, 然后以卷积系数的方式应用到所有的数据子集, 就可以计算得到每一个数据子集中心点处平滑后信号的估计值。这种方法是在发表了对应于不同多项式和子集大小的卷积系数表后而流行起来的。

除了 Savitzky-Golay 滤波器, Jade5.0 还提供了 FFT 滤波器, 允许用户在傅里叶频域空间进行图谱平滑, 具体实现过程这里就不再叙述。还有很多数据平滑的方法, 如基于小波分析的模极大值去噪算法和多区间阈值去噪算法、高斯滤波算法等。在本文中, 我们研究了基于小波分析的去噪算法和 Savitzky-Golay 滤波器, 并通过仿真模拟, 对比分析了它们对植物类和矿物类 XRD 图谱的去噪效果, 详情见文献[11], 这里我们仅给出一些主要研究结果。我们分别应用上述三种去噪算法对信噪比为 30dB 的 SiO_2 和藏药石伟 XRD 图谱进行了去噪处理, 见图 2-5 和 2-6。MJ、Thr 和 SG 分别代表模极大值去噪算法、多区间阈值去噪算法和 Savitzky-Golay 滤波器。

从图 2-5 中, 我们可以看出, 无论应用哪种算法, SiO_2 图谱中的噪声几乎消除干净, 但效果还是有些不同。SG 算法的最好结果表现出很好的去噪效果, 峰型并没有明显失真, 但也导致了一些统计起伏; 还有, 尽管 MJ 算法能很好的平滑图谱, 不像 SG 那样带来一些数据统计起伏, 但它却不能分辨间距较小的双峰, 从而导致信号失真; 至于 Thr 算法, 其去噪效果明显很好, 但却不能很好的保留峰型。同样地, 图 2-6 表明, 和原始图谱相比, MJ 和 SG 算法基本上达到了

去除噪声的目的，但在衍射角 16° 处的两个重叠峰消失了，取而代之的是一个宽峰；还出现了边界效应，见图中下三角形标记处，对于 MJ 来说，可以通过修改程序加以避免。对于 Thr 算法处理后的石伟 XRD 图谱，一些峰明显消失，同时衍射峰强度降低了，见图中上三角形、实心钻石形和下箭头标记处； 16° 处的那个峰也是如此。这些表明在处理像石伟一样的植物类 XRD 图谱时，去噪效果很差。

总的来说，对于包含较多“尖峰”的植物类 XRD 图谱，Thr 算法表现出很好的平滑效果，而 MJ 算法更加适合对包含宽弥散峰的植物类 XRD 谱进行去噪处理。尽管对于植物类和矿物类 XRD 图谱，最佳的 SG 结果表现出比较好的去噪特性，但最佳参数的确定较难，需要人工指定。

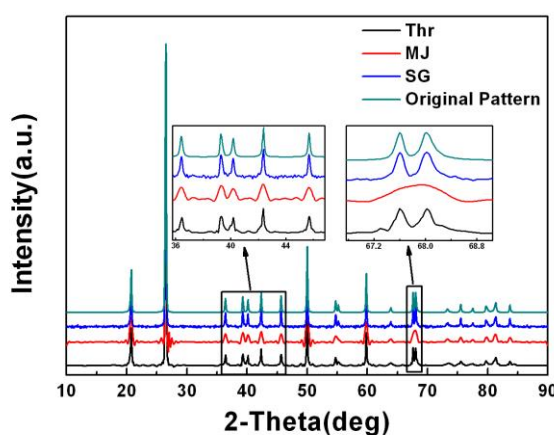


图 2-5 MJ, Thr 和 SG 去噪算法处理后的 SiO_2 图谱

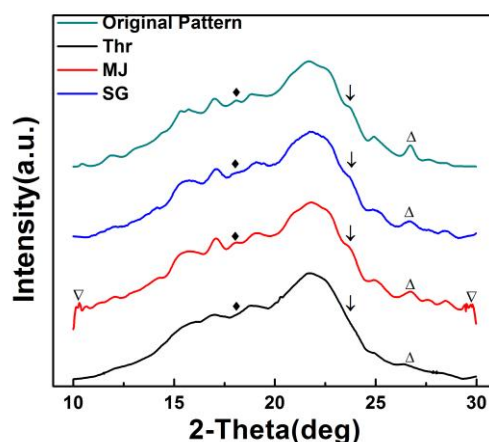


图 2-6 MJ, Thr 和 SG 去噪算法处理后的石伟图谱

2.3 背景及 $K\alpha$ 扣除

若实验样品具有较高的结晶度，其 XRD 图谱的背景应该是很平的，仅在接近直射光的极低角度区域才快速升高；若含有无定形物质或是高度分散的晶体，XRD 图谱就会呈现一个或多个可能互相重叠的近似高斯曲线的宽弥散峰。经过去噪处理后的 XRD 图谱背景线仍能看到一些小的数值起伏，这是计数的统计起伏导致的，X 射线强度的微小变化对此也有影响。这些起伏不利于图谱背景的扣除和弱峰的辨认。在进行某些处理前，必须要作背景扣除。X 射线衍射实验一般使用 K 系辐射，其包括 $K\alpha$ 和 $K\beta$ 辐射，由于二者较大的波长差异， $K\beta$ 辐射会被仪器滤掉，剩下的就只有 $K\alpha$ 辐射。但是， $K\alpha$ 辐射中又包括两种波长差很小的 $K\alpha_1$ 和 $K\alpha_2$ 辐射，它们的强度比一般情况下刚好是 2/1。在精确计算点阵常数前必须将 $K\alpha_2$ 扣除。

下面介绍应用 Jade5.0 进行背景和 $K\alpha$ 扣除的过程。

首先打开 Jade5.0 软件, 读取一个图谱文件; 然后鼠标左击主工具栏处的背景扣除按钮一次, 主窗口图谱下就会出现一条背景线, 用户这时可以控制红点手动调整背景线, 见图 2-7; 左击背景扣除按钮两次, 背景线以下的部分就会被扣除, 见图 2-7 插图 C; 在进行背景扣除前, 用户还可以右击 BG 按钮一次, 在出现的窗口里设置相关参数, 见图 2-7 插图 B。Jade5.0 通过应用某种背景函数对图谱背景进行拟合, 确定背景线, 来实现背景扣除的, 拟合方式有线性拟合、抛物线拟合和三次样条拟合。但是因其引入先验的背景函数和许多人为因素, 可能会导致数据的失真, 而缺乏足够的可靠性。Jade5.0 采用 Rachinger 算法实现 $K\alpha_2$ 的去除, 这个过程在背景线拟合时就进行了。如果选择了“Strip $K\alpha_2$ ”选项框, 在扣除背景的同时也会扣除 $K\alpha_2$ 。对于具有低背景或者尖峰或斜峰的 XRD 图谱, Rachinger 算法就表现的不是那么好了, 这时用户可以通过峰型拟合来加以辅助。

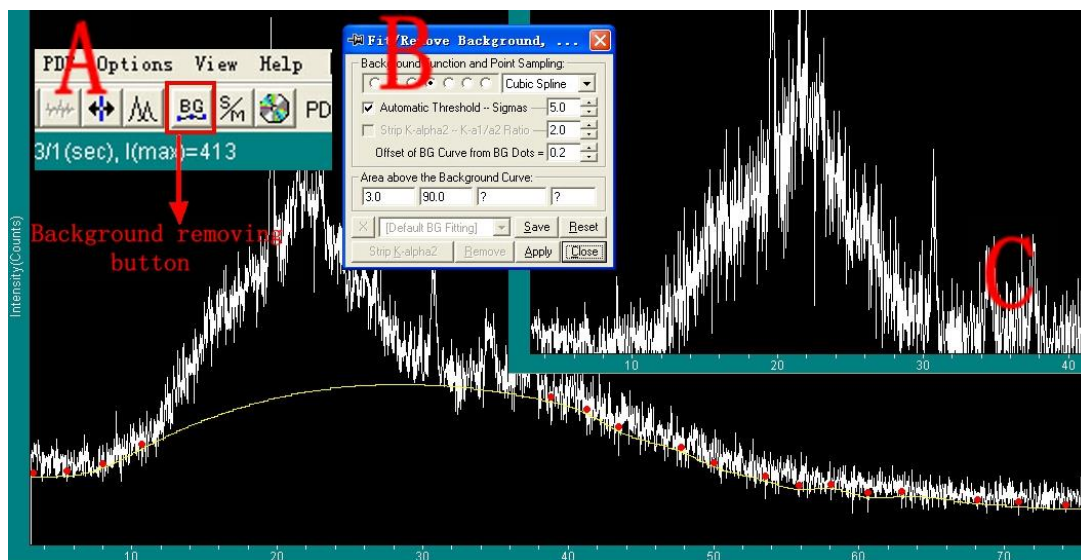


图 2-7 背景扣除操作, A:主工具栏, B:背景扣除参数设置窗口, C:左击背景扣除窗口两次后的窗口

以上是 Jade5.0 扣除背景的方法, 下面讲一下其他独立于该软件之外的算法。小波分析是基于傅里叶分析发展起来的一种时频分析工具, 广泛应用于信号处理、应用数学、物理学和地质勘探等多个领域, 非常强大。它也能用于处理图谱背景扣除问题。胡耀垓等人^[12]就充分利用小波分析的多分辨率特性, 对红外光谱信号进行多级小波分解, 保留低频系数, 置零高频系数, 接着重构恢复信号, 最后进行峰型修正, 成功实现了该类光谱信号的背景扣除处理。方勇等人^[13]也在 X 射线能谱的定量分析中引入小波分析, 进行背景扣除工作, 效果不错。还有一种算法也被广泛用于各种光谱信号的背景扣除, 叫做统计敏感的非线性迭代削峰算法 (SNIP)。如龙斌等人^[14]结合对称零面积卷积寻峰算法, 采用 SNIP 算法, 给出

了一种自适应的扣除能谱散射背景的方法。SNIP 算法具体实现可参见该文献。无论是小波分析还是 SNIP 算法，都是依据信号内在变化趋势进行背景扣除，不需要引入一些先验知识，可靠性强。

2.4 图谱寻峰

图谱寻峰是像 Jade5.0 和 Highscore 那样专业 XRD 图谱分析软件对试验样品进行物相分析的重要步骤。现在发展出来的寻峰算法很多，如对称零面积卷积法、简单比较寻峰法、二阶导数寻峰法、协方差法寻峰、线性拟合寻峰法等等。无论哪种算法都有其适用的地方，我们不能武断地说其好坏，但最好的寻峰算法却是有共同特征的：一是它具备较高的重峰分辨能力，二是能很好的甄别弱峰，三是尽可能剔除掉假峰。

下面简要介绍 Jade5.0 的寻峰过程。用户首先打开图谱文件，然后左击主工具栏中寻峰按钮(图 2-8 插图 1)，即完成寻峰处理；也可以先右击主菜单栏中的寻峰按钮，再在打开的窗口中指定相关参数，点击“Apply”按钮完成寻峰。见图 2-8。Jade5.0 是基于 Savitzky-Golay 二阶导数，并考虑到衍射强度数据的计数统计来进行寻峰处理的。用户可以在参数设置窗口选择两种“Filter Type”：Parabolic Filter(二次滤波器)和 Quartic Filter(四次滤波器)；而对于峰位的确定方式，jade 在此提供三种：峰顶，面心拟合和二次多项式拟合；其他参数的设置这里不再赘述。

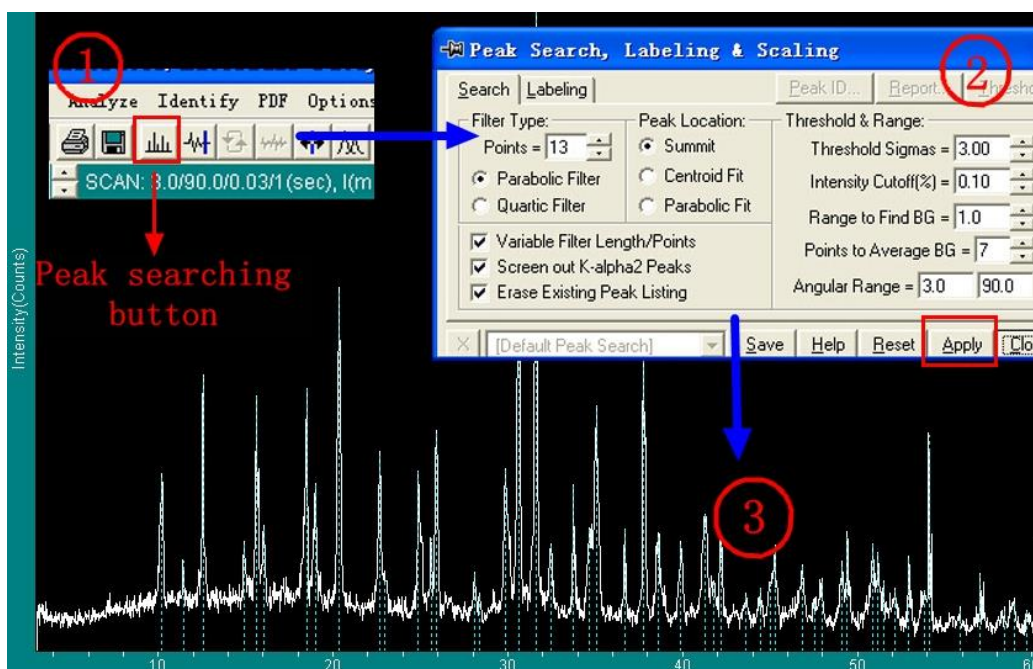


图 2-8 图谱寻峰操作 ①主工具栏 ②寻峰参数设置窗口 ③寻峰结果显示窗口

MATLAB 软件也有一个内置的寻峰算法, 名为 `findpeaks`, 其较为完整的形式是 `[pks, locs, p, w]=findpeaks(data, x)`. 这里 `data` 是等待寻峰的一组数据, 而 `x` 是相应的位置向量。返回值里, `pks` 代表数组 `data` 中所寻到的所有峰值, `locs` 代表所有峰值对应的位置, `p` 代表相应的峰宽数组, `w` 代表各峰的实际去背景高度。MATLAB 还提供了额外的输入参数用以更为精细地筛选自己所需要的峰如参数 “NPeaks” 指定函数所能寻找到的最大峰数; 参数 “MinPeakHeight” 指定函数所能寻找到峰的最小峰值; 参数 “Threshold” 指定函数所寻峰与邻峰的最小峰高差; 参数 “MinPeakDistance” 制定最小峰间距等等。若将该函数用于 XRD 图谱的寻峰处理, 那么它的形式就变为 `[PeakIntensity, PeakDegrees, Hight, FWHM]=findpeaks(Intensity, Degree)`. 这里我们应用该函数对一藏药 XRD 图谱进行寻峰, 结果见图 2-9, 具体实现过程如下:

Step1. 将原始 `mdi` 格式文件转化为只包含衍射角和衍射强度两列数据的 `txt` 文件;

Step2. 加载图谱文件, 读取衍射角 `degree`, 衍射强度 `intensity`;

Step3. 调用 `findpeaks` 函数, 寻找衍射峰,

```
[pks,locs]=findpeaks(intensity,'minpeakdistance',20,'minpeakheight',60);
```

Step4. 作图。

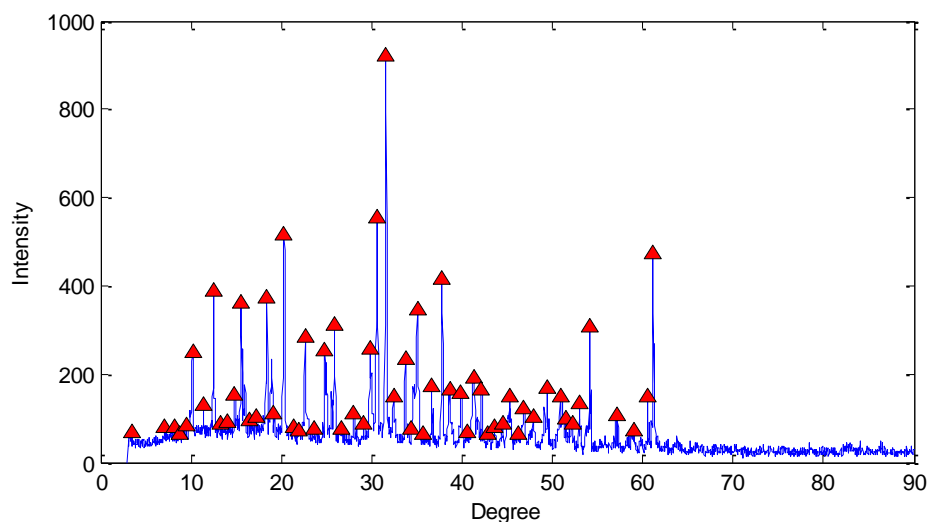


图 2-9 MATLAB 内置函数寻峰

第三章 X 射线衍射二阶导数指纹图谱

X 射线衍射(XRD)是材料分析的重要手段之一，常被用来进行物相分析，点阵常数、应力、晶粒尺寸和点阵畸变等测定工作。因为这种方法具备很多不同于其他测试手段的优点，而广泛延伸至很多领域，中药鉴定与识别^[4-7]便是其一。对于矿物类和化石类药材，它们的 XRD 图谱包含明显的特征衍射峰，峰形尖锐，峰位明确，能够非常有效地对野生或栽培和真假药材进行鉴定和分类。但是面对植物类药材时却遇到了困难，这困难来源于其特有的 XRD 图谱特征。第一，在 5~30°衍射范围内出现一个的模糊弥散型宽峰；第二，图谱中含有较多“毛刺”；第三，不多的尖锐衍射峰叠加在弥散型宽峰之上。前两者是因为，植物类药材含有的大量有机成分，它们的结晶多属于低级晶系，晶胞大小多在 7~20Å，相应主极大衍射峰多分布于掠射角为 5~30°范围内，易形成峰的叠加和款化，出现模糊性弥散峰。由于“毛刺”较多，一些衍射峰也会被淹没隐藏。这样，若继续沿用传统的 XRD 分析法，虽然 XRD 图谱包含丰富的指纹完整信息，但其却无明显的衍射峰或者根本无峰可寻。怎么办呢？引入新的数据处理技术。这种技术要能够对弥散型图谱进行解析：去除噪声影响，检测弱峰，分离重叠峰，获得谱峰数据。从这些角度出发，目前有三种方法可以选择，它们是小波分析技术、数字分峰技术和二阶导数图谱。在这一章，我会对二阶导数指纹图谱进行详细介绍，而数字分峰技术，我们会在下一章见到。

3.1 二阶导数指纹图谱的数学基础

二阶导数方法为什么能够解析植物类 XRD 弥散型图谱，给出丰富的谱峰数据呢？2006 年，杨建华等人^[15]给出了这种方法的理论依据，并进行了相似度计算，证明其可靠性。二阶导数指纹图谱的数学理论介绍如下。

根据 XRD 线形分析，每个独立的衍射峰可由轴对称分布函数来表征，这里我们选用 Pseudo-Voigt 函数，见公式 (3-1)。因此 XRD 图谱就是由不同参数的 n 个 Pseudo-Voigt 函数的叠加，表达式见公式 (3-2)。

$$pV(x) = I_0[(1-\eta)\exp(-\frac{\pi x^2}{\beta_G^2}) + \eta \frac{1}{(1 + \pi^2 x^2 / \beta_G^2)}] \quad (3-1)$$

$$I = \sum_{i=1}^n pV_i \quad (3-2)$$

其中 $\exp(-\frac{\pi x^2}{\beta_G^2})$ 为高斯函数, $\frac{1}{1+\pi^2 x^2/\beta_G^2}$ 为洛伦兹函数, η (0~1) 为洛伦兹函数所占比例。因为这些 Pseudo-Voigt 函数都满足狄义赫利条件, 而任何满足狄义赫利条件的函数都可以用三角多项式来表达, 因此叠加线形 $I(2\theta)$ 可以写成公式 (3-3), 其中 $2N$ 为 $I(2\theta)$ 有值区间角度等分的份数, A_0, A_n 和 B_n 都是函数 $I(2\theta)$ 的傅里叶系数。

$$I(2\theta) = \frac{A_0}{2} + \sum_{n=1}^{\infty} [A_n \cos(\frac{2\pi n}{2N} 2\theta) + B_n \sin(\frac{2\pi n}{2N} 2\theta)] \quad (3-3)$$

$$\begin{cases} A_0 = \frac{1}{N} \int_1^{2N+1} I(2\theta) d2\theta \\ A_n = \frac{1}{N} \int_1^{2N+1} I(2\theta) \cos(\frac{2\pi n}{2N} 2\theta) d2\theta \\ B_n = \frac{1}{N} \int_1^{2N+1} I(2\theta) \sin(\frac{2\pi n}{2N} 2\theta) d2\theta \end{cases} \quad (3-4)$$

因此, 可以根据已知函数形式的 $I(2\theta)$, 首先计算出其傅里叶系数 A_0, A_n 和 B_n , A_0, A_n 和 B_n 则为常数。对 XRD 的叠加线形 $I(2\theta)$ 求二阶导数, 即是对方程式 (3-3) 求二阶导数, 可得:

$$\frac{\partial^2 I(2\theta)}{\partial^2 (2\theta)} = -(\frac{\pi}{N})^2 \sum_{n=1}^{\infty} [A_n \cos(\frac{2\pi n}{2N} 2\theta) + B_n \sin(\frac{2\pi n}{2N} 2\theta)] \quad (3-5)$$

通过以上运算, XRD 图谱的二阶导数曲线可由取不同参数的 n 个函数叠加拟合。由于函数 (3-5) 消除了常数项和一次项的干扰, 使原图谱的重叠衍射峰分离。比较函数 (3-3) 和 (3-5), 当函数 (3-3) 某 2θ 处取极大值时, 函数 (3-5) 在此处取极小值。因此在原图谱极大值处被隐藏, 叠加的峰, 在二阶导数图谱上可取得极小值。这就为从原 XRD 图谱中分离重叠隐藏的低频信号提供了理论依据。

将原 XRD 图谱输入软件中, 求原 XRD 图谱的二阶导数。根据上述的数学理论, 二阶导数曲线中一个极小值对应原始曲线中的一个峰位, 由于原始曲线的峰位一般为正值, 二阶导数曲线的极小值则为负值, 故取曲线中的负数部分, 对曲线的负数部分求绝对值, 得到二阶导数图谱。

X 射线衍射实验采集到图谱实质上是关于衍射角 2θ 和衍射强度的一组离散数据, 要获得其二阶导数, 连续函数的微积分已不能解决, 需要借助离散数学中的数值微分。

数值微分就是对一组离散数据求取低阶导数或高阶导数近似值的数学工具。下面着重介绍三点公式、五点公式、三次样条法、三次 B 样条法、外推法等几种二阶数值微分的数学理论。

3.1.1 三点公式法

三点公式、五点公式等是数值微分常用的公式。我们可以利用插值公式进行构造，也可以对其进行 Taylor 展开，借助 Richardson 外推算法，进一步得到高精度的二阶数值微分公式。下面我们首先给出三点公式，介绍其如何应用，然后应用 Matlab 和 Mathematica 编程实现，最后再给出公式的具体推导过程。

(1) 编程实现

若函数 $f(x)$ 在节点 $x_i(i=0,1,2)$ 处的函数值 $f(x_i)$ 已知，那么相应的二阶三点公式如下：

$$\begin{cases} f''(x_0) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} \\ f''(x_1) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} \\ f''(x_2) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} \\ h = x_1 - x_0 = x_2 - x_1 \end{cases}$$

这里的 h 是在等距节点情况下的步长。依据这个公式，利用三个点即 $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$ 就可以求出任意一点处的二阶导数值，若是对于包含大量节点的图谱信号，则利用其中一个公式，从左边开始选择三点参与运算，然后依次向右移动再选择三点进行计算，直至信号结束，见下图。该运算过程相当于应用三点公式中的系数 $(1, -2, 1)/h^2$ 作为卷积核，与信号进行卷积运算。我们分别在 Matlab 和 Mathematica 软件设计平台编写了相应程序，他们能够获取藏药材 XRD 图谱的二阶导数图谱数据，如下：

Mathematica

```
data = Import["test.txt", "Table"];
ker = {-1, 16, -30, 16, -1}/(12*0.1^2);
ListLinePlot[ListConvolve[ker, data], PlotRange -> Full]
```

MATLAB

```
function sandian2_bianwei0(dir1, dir2)
fid=fopen(dir1, 'r'); %打开数据文件
fseek(fid, 25, 0);
a=importdata(dir1); %读取数据文件存入数组 a 中
x=a(:, 1); %第 1 列存到 x
```

```

fi=a(:,2); %第2列存到fi
h=x(2)-x(1);
n=length(fi);yxx=zeros(1,n);
yxx(1)=(fi(1)-2*fi(2)+fi(3))/(h^2);
for k=2:n-1
    yxx(k)=(fi(k-1)-2*fi(k)+fi(k+1))/(h^2);
end
yxx(n)=(fi(n-2)-2*fi(n-1)+fi(n))/(h^2);
for k=1:n
    if yxx(k)>0
        yxx(k)=0;
    end
end
plot(x,abs(yxx),'k');
title('sandian2_bianwei0');
ylabel('Intensity(counts)');
xlabel('2-theta');
b(1,:)=x;b(2,:)=abs(yxx);
file = fopen(dir2,'wt');
fprintf(file, '%6.2f      %12.8f\n', b);
fclose(file);
disp('success!!!');

```

(2) 公式推导

在文献[16]中,夏爱生等人详细给出了二阶三点数值微分公式的推导过程,还利用 Richardson 外推算法,提高其收敛阶数,得到高精度的二阶数值微分公式,在这里我们不会加以赘述,只给出另一个较为简单的推理过程,即利用插值公式进行构造。

若函数 $f(x)$ 在节点 $x_i(i=0,1,2)$ 处的函数值 $f(x_i)$ 已知,那么就可以构造 $f(x)$ 的 n 次插值多项式 $P_n(x)$,随后可用其导数近似代替被插值函数 $f(x)$ 的导数。该插值多项式如下:

$$\begin{cases} P_n(x) = \sum_{i=0}^n y_i l_i(x) \\ l_i(x) = \frac{\omega(x)}{(x-x_i)\omega'(x_i)}, \quad \omega(x) = (x-x_0)\cdots(x-x_n) \\ ERROR: R(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \xi \in (a, b) \end{cases}$$

考虑 $n=2$ 的情况,已知节点及相应函数值, $x_0, x_1=x_0+h, x_2=x_0+2h, y_0=f(x_0), y_1=f(x_1), y_2=f(x_2)$. 设满足上述插值条件的二次多项式为 $P_2(x)$,

$$\begin{aligned}
 P_2(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} y_2 \\
 &= \frac{(x-x_1)(x-x_2)}{2h^2} y_0 + \frac{(x-x_0)(x-x_2)}{-h^2} y_1 + \frac{(x-x_0)(x-x_1)}{2h^2} y_2
 \end{aligned}$$

则

$$P'_2(x) = \frac{2x-x_1-x_2}{2h^2} y_0 + \frac{2x-x_0-x_2}{-h^2} y_1 + \frac{2x-x_0-x_1}{2h^2} y_2$$

于是可得一阶三点公式

$$\begin{cases}
 f'(x_0) \approx P'_2(x_0) = \frac{1}{2h} [-3f(x_0) + 4f(x_1) - f(x_2)] \\
 f'(x_1) \approx P'_2(x_1) = \frac{1}{2h} [-f(x_0) + f(x_2)] \\
 f'(x_2) \approx P'_2(x_2) = \frac{1}{2h} [f(x_0) - 4f(x_1) + 3f(x_2)]
 \end{cases}$$

其相应的余项

$$\begin{cases}
 f'(x_0) - P'_2(x_0) = \frac{h^2}{3} f'''(\xi_0) \\
 f'(x_1) - P'_2(x_1) = -\frac{h^2}{6} f'''(\xi_1), \quad (\xi_0, \xi_1, \xi_2 \in [x_0, x_2]) \\
 f'(x_2) - P'_2(x_2) = \frac{h^2}{3} f'''(\xi_2)
 \end{cases}$$

对 $P'_2(x)$ 再求导, 得 $P''_2(x) = \frac{1}{h^2} [f(x_0) - 2f(x_1) + f(x_2)]$, 这样就建立了计算二阶导数近似值的数值微分公式

$$\begin{cases}
 f''(x_0) \approx P''_2(x_0) = \frac{1}{h^2} [f(x_0) - 2f(x_1) + f(x_2)] \\
 f''(x_1) \approx P''_2(x_1) = \frac{1}{h^2} [f(x_0) - 2f(x_1) + f(x_2)] \\
 f''(x_2) \approx P''_2(x_2) = \frac{1}{h^2} [f(x_0) - 2f(x_1) + f(x_2)]
 \end{cases}$$

其对应的余项为

$$\begin{cases} f''(x_0) - P_2''(x_0) = -hf''(\xi_1) + \frac{h^2}{6} f^{(4)}(\xi_2) \\ f''(x_1) - P_2''(x_1) = -\frac{h^2}{12} f^{(4)}(\xi_3) \\ f''(x_2) - P_2''(x_2) = hf''(\xi_4) + \frac{h^2}{6} f^{(4)}(\xi_5) \end{cases}$$

3.1.2 五点公式法

首先给出二阶五点公式的具体形式。若 $f(x), x \in [a, b]$ 在等距节点 $a \leq x_0 < x_1 < x_2 < x_3 < x_4 \leq b$ 处的函数值 $f(x_k), (k=0,1,2,3,4)$ 已知, 且 $x_{k+1} - x_k = h$, 那么节点 $x_k (k=0,1,2,3,4)$ 处的二阶导数五点数值微分公式如下:

$$f''(x_0) = \frac{1}{12h^2} [35f(x_0) - 104f(x_1) + 114f(x_2) - 56f(x_3) + 11f(x_4)] \quad (3-6)$$

$$f''(x_1) = \frac{1}{12h^2} [11f(x_0) - 20f(x_1) + 6f(x_2) + 4f(x_3) - f(x_4)] \quad (3-7)$$

$$f''(x_2) = \frac{1}{12h^2} [-f(x_0) + 16f(x_1) - 30f(x_2) + 16f(x_3) - f(x_4)] \quad (3-8)$$

$$f''(x_3) = \frac{1}{12h^2} [-f(x_0) + 4f(x_1) + 6f(x_2) + 20f(x_3) + 11f(x_4)] \quad (3-9)$$

$$f''(x_4) = \frac{1}{12h^2} [11f(x_0) - 56f(x_1) + 114f(x_2) - 104f(x_3) + 35f(x_4)] \quad (3-10)$$

像三点公式一样, 我们可以利用插值公式来构造五点数值微分公式, 简单的过程就是, 在区间 $[a, b]$ 上作 $f(x)$ 的 4 次 Lagrange 插值函数, 将 $x = x_0 + th, t \in [0, 4], x_k = x_0 + kh$ 代入, 并将方程两端对 t 求二次导数, 再分别把 $t=0, 1, 2, 3, 4$ 代入, 即可得到 $x_k (k=0,1,2,3,4)$ 节点二阶导数的 5 点数值微分公式。五点公式具体应用和三点公式一样, 都可以用移动窗口加权求和的思想来理解。

(1) 编程实现

基于上述的 5 个二阶五点公式, 运用 MATLAB 和 MATHEMATICA 语言进行编程。在编程过程中, 有两点问题需要说明。一是可以选择任何一个公式或多个公式进行编程。由于 XRD 图谱信号的数据点很大, 首尾两处端点的导数可以舍弃不予计算, 对结果影响极小, 当然若选择相应的公式对端点计算就可以获得全部数据点的二阶导数。二是, 运算结果出来后会现正的或负的数据点。根据 3.1 节中的分析, 我们知道只有负的数据点对我们才是有意义的。我们需要对正值进行置零处理, 然后取其绝对值。考虑这两点, 编写的程序如下:

Mathematica

```
data = Import["test.txt", "Table"];
ker = {-1, 16, -30, 16, -1}/(12*0.1^2);
ListLinePlot[ListConvolve[ker, data], PlotRange -> Full]
```

MATLAB

```
function wudian2_bianwei0(dir1,dir2)
%函数名: wudian2
%功能: 利用二阶导数的五点公式求算离散点处的二阶微商
%h 为步长
%返回值为离散点处的近似二阶微商 yxx, 以及离散点的集合
fid=fopen(dir1,'r'); %打开数据文件
fseek(fid,25,0);
a=importdata(dir1); %读取数据文件存入数组 a 中
x=a(:,1); %第 1 列存到 x
fi=a(:,2); %第 2 列存到 fi
h=x(2)-x(1);
n=length(fi); yxx=zeros(1,n);
yxx(1)=(fi(1)-2*fi(2)+fi(3))/(h.^2);
yxx(2)=(fi(1)-2*fi(2)+fi(3))/(h.^2);
for k=3:n-2
    yxx(k)=(-fi(k-2)+16*fi(k-1)-30*fi(k)+16*fi(k+1)-fi(k+2))/(12.*h.^2);
end
yxx(n-1)=(fi(n-2)-2*fi(n-1)+fi(n))/(h.^2);
yxx(n)=(fi(n-2)-2*fi(n-1)+fi(n))/(h.^2);
fclose(fid); %文件关闭
for k=1:n
    if yxx(k)>0
        yxx(k)=0;
    end
end
plot(x,abs(yxx),'k');
title('wudian2_bianwei0');
ylabel('Intensity(counts)');
xlabel('2-theta');
b(1,:)=x;b(2,:)=abs(yxx);
file = fopen(dir2,'wt');
fprintf(file, '%6.2f %12.8f\n', b);
fclose(file);
disp('success!!!');
```

(2) 公式推导

有关二阶导数五点微分公式的外推算法，在文献[17]中已进行了详细推理，在这里我们只简要给出中点节点五点微分公式的外推算法，如下。

分别将 $f(x_0)$, $f(x_1)$, $f(x_2)$, $f(x_3)$, $f(x_4)$ 在 x_2 点作 Taylor 展开并代入 (3-8) 式的右端整理得到

$$f''(x_2) - S_{2,1}(h) = a_1 h^4 + a_2 h^6 + a_3 h^8 + \dots \quad (3-11)$$

式中： $S_{i,k+1}$ 表示 x_i 节点处第 k 次外推公式

$$S_{2,1}(h) = \frac{-f(x_0) + 16f(x_1) - 30f(x_2) + 16f(x_3) - f(x_4)}{12h^2}$$

$$\alpha_1 \frac{f^{(5)}(x_2)}{90}, \alpha_2 \frac{f^{(7)}(x_2)}{1008}, \alpha_3 \frac{f^{(9)}(x_2)}{21600},$$

对于固定的 x_2 , $\alpha_i (i=1, 2, \dots)$ 是与 h 无关的常数，所以上面的误差估计式符合 Richardson 外推算法，将 h 缩小一倍得到，

$$f''(x_2) - S_{2,1}\left(\frac{h}{2}\right) = a_1 \left(\frac{h}{2}\right)^4 + a_1 \left(\frac{h}{2}\right)^6 + a_1 \left(\frac{h}{2}\right)^8 + \dots \quad (3-12)$$

由 $16 \times (3-12) - (3-11)$ 式整理得到

$$f''(x_2) - S_{2,2}(h) = a_1 \left(\frac{h}{2}\right)^4 + a_1 \left(\frac{h}{2}\right)^6 + a_1 \left(\frac{h}{2}\right)^8 + \dots$$

$$\text{式中: } \alpha_2' = -\frac{1}{20}\alpha_2, \alpha_3' = -\frac{1}{16}\alpha_3, S_{2,2}(h) = \frac{16S_{2,1}\left(\frac{h}{2}\right) - S_{2,1}\left(\frac{h}{2}\right)}{15}$$

由此可见， x_2 点的五点微分公式外推一次后，精度由 Oh^4 提高到 Oh^6 。

依次类推，可得到外推算法的递推序列：

$$\left\{ \begin{array}{l} S_{2,1}(h) = \frac{-f(x_0) + 16f(x_1) - 30f(x_2) + 16f(x_3) - f(x_4)}{12h^2} \\ S_{2,k+1}(h) = \frac{2^{2(k+1)} S_{2,k}\left(\frac{h}{2}\right) - S_{2,k}(h)}{2^{k+2} - 1}, (k=1, 2, \dots) \end{array} \right. \quad (3-13)$$

式中 k 为外推次数， $S_{2,k+1}(h)$ 的截断误差为 $O(h^{2(k+1)})$ ，既每外推一次收敛阶增长两阶。

3.1.3 三次样条法

设已知 $f(x)$ 在节点 $x_k (k=0, 1, \dots, n)$ 的函数值，利用所给数据获得相应插值函数 $S(x)$ ，并取 $S''(x)$ 的值作为 $f''(x)$ 的近似值，即

$$f''(x) \approx S''(x)$$

这就是利用插值函数求取离散数据二阶导数的方法。所谓插值即在离散数据的基础上补插连续函数，使得这条连续曲线通过全部给定的离散数据点，目前常用的插值函数类型有多项式、埃尔米特、分段和三角函数。这里介绍三次样条插值函数还有下一节的三次 B 样条插值函数。

(1) 编程应用

应用三次样条插值函数获取 X 射线衍射图谱的二阶导数图谱，有两大关键步骤，一是根据离散的图谱数据构造三次样条插值函数，二是对三次样条函数求二阶导数。下面给出具体算法步骤：

Step1. XRD 图谱文件经格式转化后，分离出衍射角度，存于 degree, 和相应的衍射强度，存于 intensity;

Step2. 构造三次样条插值函数。这里利用 MATLAB 内置函数 $S = \text{csapi}(\text{degree}, \text{intensity})$ 来进行构造，S 就是与图谱数据相对应的插值函数；

Step3. 计算二阶导数 S'' 。这里采用 MATLAB 内置函数 $S'' = \text{fnder}(S, 2)$ 。

Step4. 计算衍射角 degree 处的二阶导数值。

Step5. 将正的二阶导数值置零，负的取绝对值。

基于 MATLAB 编程语言，编写程序如下：

MATLAB

```
function yangtiaofa_bianwei0(dir1,dir2)
%三次样条插值
fid=fopen(dir1,'r'); %打开数据文件
fseek(fid,25,0);
a=importdata(dir1); %读取数据文件存入数组 a 中
xi=a(:,1); %第 1 列存到 x
fi=a(:,2); %第 2 列存到 fi
n=length(fi);
S=csapi(xi,fi);
dsp=fnder(S,2);
y=fnval(dsp,xi);
for k=1:n
    if y(k)>0
        y(k)=0;
    end
end
plot(xi,abs(y),'k');
title('yangtiaofa_bianwei0');
ylabel('Intensity(counts)');
xlabel('2-theta');
b(1,:)=xi;b(2,:)=abs(y);
file = fopen(dir2,'wt');
```

```
fprintf(file, '%6.2f      %12.8f\n', b);
fclose(file);
disp('success!!!');
```

(2) 三次样条函数的推导过程

设 $[a, b]$ 上有插值节点 $a = x_1 < x_2 < \cdots < x_n = b$ ，对应函数值为 y_1, y_2, \dots, y_n 。若函数 $S(x)$ 满足 $S(x_j) = y_j$ ($j = 1, 2, \dots, n$)， $S(x)$ 在 $[x_j, x_{j+1}]$ ($j = 1, 2, \dots, n-1$) 上都是不高于三次的多项式。当 $S(x)$ 在 $[a, b]$ 具有二阶连续导数，则称 $S(x)$ 为三次样条插值函数。要求 $S(x)$ 只需在每个子区间 $[x_j, x_{j+1}]$ 上确定一个三次多项式，设为

$$S_j(x) = a_j x^3 + b_j x^2 + c_j x + d_j, (j = 1, 2, \dots, n-1) \quad (3-14)$$

其中 a_j, b_j, c_j, d_j 待定，并要使它满足

$$\begin{cases} S(x_j) = y_j, S(x_j - 0) = S(x_j + 0), (j = 2, \dots, n-1) \\ S'(x_j - 0) = S'(x_j + 0), S''(x_j - 0) = S''(x_j + 0), (j = 2, \dots, n-1) \end{cases} \quad (3-15)$$

式子(3-14)、(3-15)共给出 $n+3(n-2)=4n-6$ 个条件，需要待定 $4(n-1)$ 个系数，因此要唯一确定三次插值函数，还要附加 2 个边界条件。通常由实际问题对三次样条插值在端点的状态要求给出。文献[18]中，徐小勇等人以第一类边界条件为例，用节点处二阶导数表示三次样条插值函数，用追赶法求解相关方程组，构造出了三次样条插值函数。对于具体的推导过程，这里不再赘述，详细请参看该文献。

3.1.4 三次 B 样条法

应用三次 B 样条函数对 XRD 数据进行插值处理以获得其二阶导数图谱的步骤和上一节介绍的三次样条法相差不大。上一节中，步骤 2 是利用 MATLAB 的内置函数 $S = \text{csapi}(\text{degree}, \text{intensity})$ 构造三次样条插值函数，这里只需将其替换成 $S = \text{spapi}(k, \text{degree}, \text{intensity})$ 就可以了。 k 是一个正整数，仅用于指定想要的样条阶数，我们选择 $k=3$ ，即三次 B 样条函数。 degree 代表衍射角度 2θ ， intensity 代表相应的衍射强度。下面给出其具体程序实现，至于 B 样条插值函数构造的过程，这里就不加以叙述了，可参考文献[19-20]

```
function yangtiaofaB_bianwei0(dir1, dir2)
%B 样条插值（远优于三次样条插值）
fid=fopen(dir1, 'r'); %打开数据文件
fseek(fid, 25, 0);
a=importdata(dir1); %读取数据文件存入数组 a 中
```

```

xi=a(:,1);    %第1列存到x
fi=a(:,2);    %第2列存到fi
n=length(fi);
S=spapi(5,xi,fi);
dsp=fnder(S,2);
y=fnval(dsp,xi);
for k=1:n
    if y(k)>0
        y(k)=0;
    end
end
plot(xi,abs(y),'k');
title('yangtiaofaB_bianwei0');
ylabel('Intensity(counts)');
xlabel('2-theta');
b(1,:)=xi;b(2,:)=abs(y);
file = fopen(dir2,'wt');
fprintf(file, '%6.2f      %12.8f\n', b);
fclose(file);
disp('success!!!');

```

3.2 植物类药材粉末衍射文件

3.2.1 获取 PDF 文件的方法

上面的 3.1 节中，我们对二阶导数图谱理论进行了原理上的介绍，给出了四种二阶数值微分方法的简单推理过程，还基于 MATLAB 语言编写了相应的程序。在这些程序的基础上，在这一小节，我们提出一种获取植物类药材粉末衍射文件（PDF）的处理流程。通过该流程，我们不仅能够获取藏药材的二阶导数指纹图谱，还能建立相应的粉末衍射文件。所谓 PDF（Powder Diffraction File），是国际衍射数据中心编辑、发行的纯化何物 X 射线衍射数据集。该数据集以表格形式向使用者提供化合物分子式、英文名称、晶系、空间群、晶面间距、相对强度、晶胞参数和 Miller 指数等内容。这些数据是进行物相检索和分析的基础。该数据库中药材类 PDF 极为有限，大大限制了 XRD 技术在中药质量控制和鉴定方面的应用。因此，这里尝试建立获取藏药材 PDF 的处理流程，对藏药材的深入研究具有一定的意义。

获取植物类药材粉末衍射文件的过程包括如下步骤：

（1）获取植物类药材的 X 射线衍射图谱文件

将藏药材在 70℃ 的烘箱中烘干，然后利用粉碎机粉碎，再通过 200 目筛，得到其粉末状样品，最后进行 XRD 测试，得到样品的 X 射线衍射图谱文件；

(2) 获取植物类药材的二阶导数指纹图谱

首先对(1)产生的 XRD 原始文件进行格式转化即 MDI 格式转化为 TXT 格式,然后采用最小二乘移动平滑法对其进行平滑处理,再从 3.1 节介绍的三点公式、五点公式、三次样条法、三次 B 样条法、外推法等方法中选择一个,计算图谱信号的二阶导数,最后对负值部分取绝对值、正值置零,获得样品的二阶导数指纹图谱;

(3) 寻峰处理

对(2)中产生的二阶导数图谱作寻峰处理。这里我们选择对称零面积卷积法作为寻峰算法,指定其中的阈值因子为 0.1~1。

(4) 制作植物类药材的粉末衍射文件

根据国际标准的 PDF 格式,填写藏药材的基本信息,(3)中寻峰产生的衍射数据,以及根据布拉格公式获得的晶面间距数据,制作该藏药材相应的 PDF 文件。

根据以上获取植物藏药材 PDF 文件的流程,基于 MATLAB 语言,编写相应的程序,分别为 formatTrans.m,用于格式转化;secondDiff.m,用于获得 XRD 图谱的二阶导数;peakSearch.m,用于对二阶导数图谱进行寻峰,获得衍射峰值数据;xrdPic.m,用于制作 PDF 文件所需要的衍射峰图片;这些程序源代码见论文后附件。

3.2.2 应用实例

这里以藏药材五味子为例,应用 3.2.1 节提出的处理流程获得其相应的 PDF 文件。

(1) 首先对藏药材五味子进行 XRD 测试,获取原始 XRD 图谱。选择的 XRD 测试仪器为 Y-4Q 型 X 射线衍射仪,参数见表 3-1。

表 3-1. XRD 测试条件

管压	管流	K α	滤波片	扫描方式	扫描范围	步宽
35Kv	20mA	Cu	Ni	连续扫描	10° -80°	0.03°

(2) 然后应用编写的 MATLAB 程序对原始 XRD 图谱进行处理。图 3-1 是五味子的原始 XRD 图谱,从中我们可以看出,图谱毛刺很多,无明显衍射峰,整体呈模糊弥散型,这是由于纤维、淀粉、蛋白质、多糖等非晶态大分子有机物质造成的。图 3-2 和 3-3 分别是程序处理得到的二阶导数图谱和衍射峰图。我们发现,与原始 XRD 图谱相比,处理后图谱峰数较多,峰形尖锐,峰位明确,能够充分体现五味子 XRD 图谱的特征,反映出其整体信息。

(3) 整理藏药材五味子的编号、中文名称、拉丁文、产地等基本信息,连同(2)获得的五味子衍射峰数据,根据国际 PDF 文件标准,填入特别制作的文件表格中,

制作五味子 PDF 文件，见表 3-2。

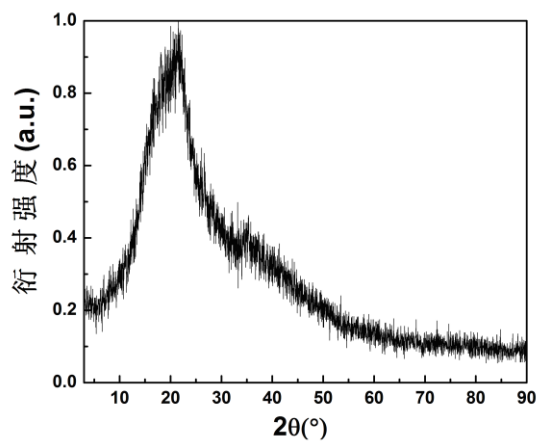


图 3-1 五味子的 XRD 原始图谱

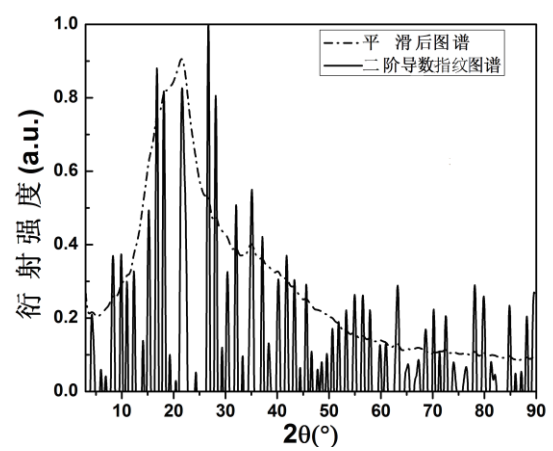


图 3-2 五味子的二阶导数指纹图谱

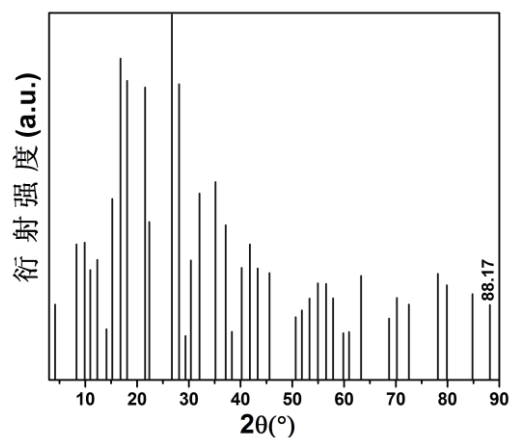
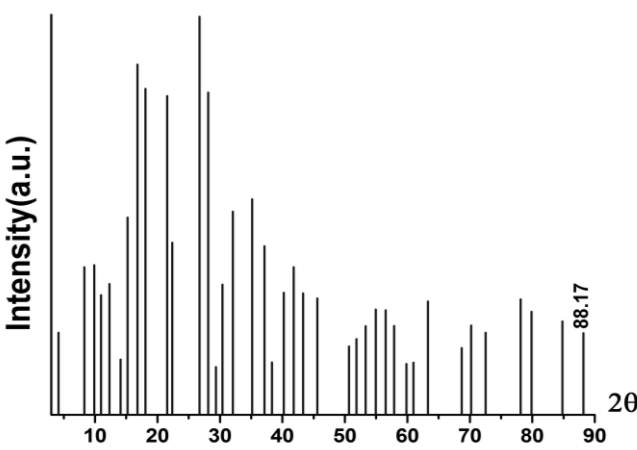


图 3-3 五味子的 XRD 衍射角度与强度

表 3-2 五味子的 PDF 卡片

d	3.3350	5.2771	4.9036	3.1710	Medicine Information	
					编号: 2011048gh	
					中文名称: 五味子	
I/I ₀	100.0	87.537	81.495	80.566	拉丁文: Schisandra chinensls (Turcz) Bail	
Experimental condition:						
SS/FOM:						
I/I _{cor} :						
Rad:	Cu-Kα					
Lambda:	1.54178					
Filter:	Ni					
d-sp:						
Molecular Weight:						
Volume[CD]:						
Dx:						
Dm:						
Sys:			2-Theta	d	I%	hkl
Lattice:			8.31	10.64	36.946	
S.G.:			9.87	8.9612	37.389	
Cell Parameters:			10.98	8.0576	29.92	
a	b	c	12.33	7.1783	32.721	
α	β	γ	15.21	5.8249	49.338	
Remarks: Pattern data was obtained at room temperatrue. This medcine was collected in Hezuo, Gansu Province, China, which belongs Magnoliaceae, growing in humus or sandy loam.			16.8	5.2771	87.537	
			18.09	4.9036	81.495	
			21.54	4.1253	79.663	
			22.38	3.9724	43.054	
			26.73	3.335	100	
			28.14	3.171	80.566	
			30.39	2.9411	32.576	
			32.07	2.7908	50.776	
			35.16	2.5523	53.884	
			37.14	2.4207	42.142	
			40.23	2.2416	30.531	
			41.79	2.1614	36.894	
			43.32	2.0886	30.373	
			45.6	1.9893	29.132	
			54.96	1.6706	26.335	
			56.52	1.6282	26.157	
			63.3	1.4691	28.341	
			78.15	1.223	28.886	
			79.89	1.2007	25.77	
			84.84	1.1428	23.391	

第四章 藏药 XRD 全谱分峰系统

4.1 章节引言

将 X 射线衍射法应用于中药测量和研究起始于 90 年代, 主要应用于矿物类和化石类药材, 其目的是从某些中药材中确定出某些无机物相成分, 并列出相应 XRD 图谱中的峰位和相对强度, 从而间接达到某些中药的目的。尽管 X 射线衍射指纹图谱的研究已有报道, 但中药 X 射线衍射指纹图谱的研究仍处于初始阶段。对于藏药研究亦是如此, 虽然国内外利用 XRD 法进行藏药指纹图谱的研究报道有一些, 但研究局限于藏药材的定性识别, 沿用了材料科学中的“物相分析”方法, 通过识别衍射图谱中衍射峰的归属来鉴别藏药。所谓的 X 射线衍射 Fourier 谱分析法实际上只是一个普通的 XRD 谱图测量, 对图谱进行平滑处理, 利用仪器缺省值简单寻峰的过程, 并没有对 XRD 谱图进行分峰计算以得到真正的衍射峰和结晶度等参数。这种传统的 XRD 数据处理模式, 对于某些矿物类和化石类中药较为有效, 而对于植物类药材, 因其所含大量有机成分的结晶多属低级晶系, 晶胞大小多在 $7\sim 20\text{\AA}$, 相应主极大衍射峰多分布于掠射角为 $5\sim 30^\circ$ 范围内, 易形成峰的叠加和款化, 出现模糊性弥散峰(见图 4-1, 2), 致使其 XRD 指纹图谱无明显特征峰或根本无峰可寻则存在致命缺陷。

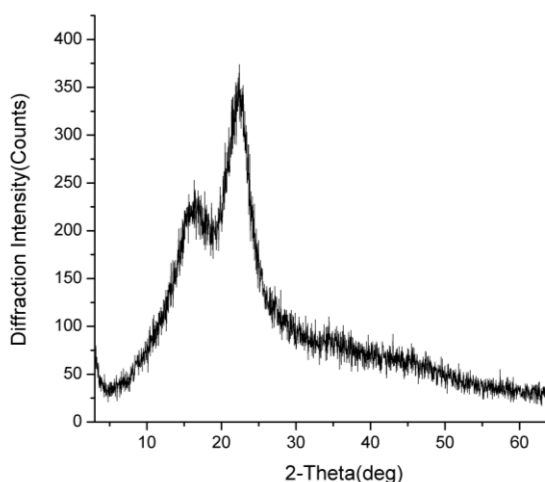


图 4-1 藏药悬钩木的 XRD 图谱

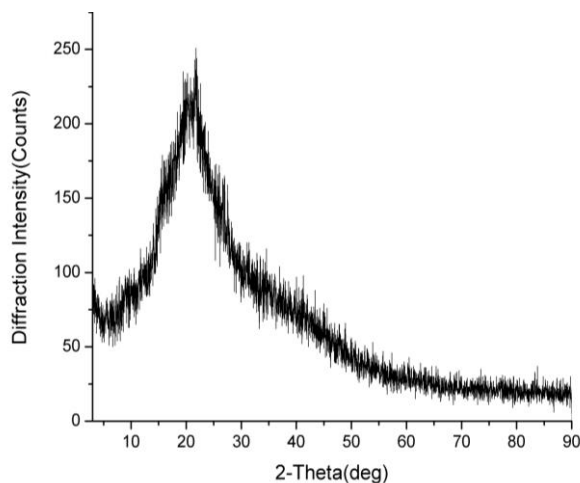


图 4-2 藏药甘青青兰的 XRD 图谱

为解决此类植物类模糊弥散 XRD 图谱解析问题, 本专著在上面章节中也提出了基于二阶导数数值微分获取藏药 XRD 图谱对应的 SD 图谱的理论, 结果表明其能够有效地将一些晶态化学成分各自专属的衍射锐峰显示出来, 挖掘出原始图谱中隐藏的信息。这里介绍另一种解决方法。

本章节基于传统的 X 射线衍射数据处理模式，即首先进行图谱平滑去噪、扣除背景，然后直接寻峰的过程，提出在在对图谱整体轮廓识别的同时，引入图谱分峰和全谱数字化拟合技术，采用 Pseudo-Voigt 函数模型进行全谱数字化拟合分峰，获得相应峰位和相对强度的新的 XRD 图谱数字分峰处理模式，以实现植物类弥散 XRD 图谱的解析。模式流程见图 4-3。

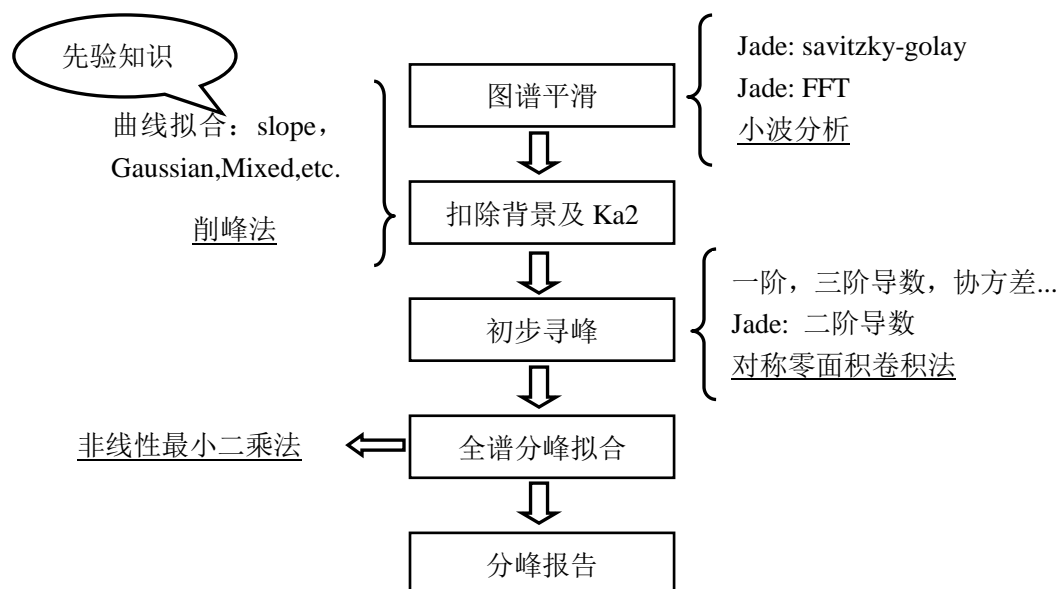


图 4-3 XRD 图谱数字分峰处理流程

格式转化模块是针对 XRD 图谱分析软件 JADE 特有的 MDI 文件格式而设计，将其转化为 TXT 文本文件，便于其他功能模块和其他图谱分析软件的应用。平滑滤波模块用于消除 X 射线衍射强度采集过程的统计起伏而引起的噪声影响，这里采用基于小波分析的模极大值去噪技术。现有的图谱分析软件 JADE 采用 Savitzky-Golay 滤波器进行平滑处理，但其不能依据图谱本身自适应地确定滤波窗口大小和曲线拟合阶数，需要手工调整。而小波模极大值去噪因其能对信号多尺度分析和依据信号和噪声各尺度间不同的传播特性准确去掉噪声，更具优势。背景扣除模块根据背景分布趋势进行削峰计算，最终确定背景并扣除之，实现 XRD 图谱中的基线校正，扣除背景。初步寻峰模块采用对称零面积卷积寻峰算法对 XRD 图谱进行寻峰处理，这里可以获得精确的衍射峰峰位，和初步估计的衍射峰强作为下个模块曲线拟合初值。现有的 JADE 软件采用二阶导数对图谱寻峰，虽然此种方法能够分辨弱峰和重峰，但带来较大的数据统计起伏，不利于后期分峰拟合运算。分峰拟合模块以高斯函数和柯西函数的混合为衍射峰剖面函数形式，采用最小二乘法对扣除背景后的平滑图谱进行全谱拟合，获得 X 射线衍射图谱的数字分峰拟合报告。报告包含衍射峰峰位、峰强、半高宽、绝对强度、相对

强度、峰型因子和偏态因子。下面章节将对本图谱分峰系统涉及到的的小波去噪、背景扣除、对称零面积卷积寻峰和分峰拟合理论进行原理上的详细介绍。

4.2 XRD 图谱数字分峰理论体系

4.2.1 小波去噪

Savitzky-Golay 法^[21]在图谱平滑模块中最为常用，如专业的 XRD 分析软件 JADE，但因其需要人为确定窗口宽度和多项式阶数，以及植物类 XRD 图谱弥散型的特点，导致平滑后峰形失真过大，所以有必要寻找新的信号处理方法。小波分析，又称为多分辨率分析，是从傅里叶分析发展而来的新时频分析工具，在时域和频域同时具有良好的局部化特性，常被称为信号分析的“数学显微镜”。近年来，小波分析的理论和方法在很多领域得到了广泛的应用。在信号处理领域人们应用小波进行去噪，并获得了非常好的效果，这得益于小波变换具有低熵性、多分辨率、去相关性和选基灵活性等多种突出特点。发展至今，有很多小波去噪方法被提出与发展，基本方法如模极大值重构^[22,23,24]、空域相关去噪^[25]和小波域阈值去噪^[26]，但总体上都包括三个基本的步骤即：对含噪声信号进行小波变换；对变换得到的小波系数进行某种处理，以去除其中包含的噪声；对处理后的小波系数进行小波逆变换，得到去噪后的信号。小波去噪方法的不同之处集中在第二步。在这里，我们选用小波模极大值重构方法对 XRD 图谱进行平滑处理。模极大值法是根据信号和噪声在多尺度空间上小波系数的模极大值传播规律的不同而发展起来的一种去噪算法。原则上只要信号与噪声的奇异性有差异，就能产生很好的去噪效果。文献[22]对二进小波变换原理、信号和噪声在小波尺度上的传播特性进行了原理上的介绍，还提出了较为详细的算法步骤。基于 MATLAB 软件设计平台，我们据此编写了图谱平滑模块的程序。

4.2.2 背景扣除

若实验样品具有较高的结晶度，其 XRD 图谱的背景应该是很平的，仅在接近直射光的极低角度区域才快速升高；若含有无定形物质或是高度分散的晶体，XRD 图谱就会呈现一个或多个可能互相重叠的近似高斯曲线的宽弥散峰。经过去噪处理后的 XRD 图谱背景线仍能看到一些小的数值起伏，这是计数的统计起伏导致的，X 射线强度的微小变化对此也有影响。这些起伏不利于图谱背景的扣除和弱峰的辨认。在进行某些处理前，必须要作背景扣除。确定衍射图的背景线，现在只有一些约定的方法。下面介绍一种“削峰法”。这种方法基于 XRD 图谱中变化迅速的特征，通过比较和它附近衍射角信息的一种方法，其显著优点是没有使用

精确的数学模型，避免了采用不同的数学模型而带来的误差。具体算法是：通过比较 $I_{2\theta_i}$ 和两个邻近衍射角上的强度值的平均值 I_m 的大小， $I_m = (I_{2\theta_{i-1}} + I_{2\theta_{i+1}})/2$ ，若 $I_{2\theta_i} > I_m$ ，衍射强度 $I_{2\theta_i}$ 替换为平均值 I_m ，在所有衍射角上依次运算一遍。当曲率比较大时，用 $I_m + C$ 代替 I_m ， C 是一个常数。可以看到，峰位置的峰幅度降低，图谱衍射峰位保持不变，经过多次循环后，剥离收敛，留下光滑的背景。不同的谱线循环次数不同，这取决于峰宽。

4.2.3 对称零面积卷积寻峰

在图谱分析领域，众多的寻峰算法被提出和应用，如简单比较法、导数法、协方差法、对称零面积卷积法(SZAC)、线性拟合法等等^[27]。他们各有优缺点，具体选择何种算法要视所研究对象的具体情况而定。本文设计的针对弥散型 XRD 图谱分峰软件要求初步寻峰模块所选算法具备较高的重峰分离能力和能够获取精确的峰位信息，而 SZAC 算法在这两方面表现不错，所以会加以优先选择。所谓的对称零面积变换寻峰是用面积为零的对称窗函数与实验测得的图谱数据进行卷积变换，对变换后的数据进行阈值处理获得衍射峰位置的方法。文献[28]对 SZAC 算法原理进行了较为详细的介绍，我们以此为基础，进行移植应用，提出其在初步寻峰模块中具体算法步骤，如下：

Step1. 按照公式(4-1)构建对称零面积变换函数，其中类峰型函数 $f(j)$ 选择高斯线型(4)，变换窗口宽度 $W=2m+1=21$ ，半高全宽 $H_G=5$ 。

$$C(j) = f(j) - \frac{1}{W} \sum_{-m}^m f(j) \quad (4-1)$$

$$f(j) = e^{-4 \ln \left[2 \left(\frac{j}{H_G} \right)^2 \right]} \quad (4-2)$$

Step2. 按照公式(4-3)对原始 XRD 图谱每个数据 y_i 进行对称零面积变换，其中 y'_i 为对应的变换谱值。

$$y'_i = \sum_{j=-m}^m C_j y_{i+j} \quad (4-3)$$

Step3. 按照公式(4-4)计算阈值函数 $T(i)$ ，然后与设定的寻峰阈值 T_0 进行比较，若在某位置 i_{peak} 处 $T(i_{\text{peak}})$ 取值大于 T_0 ，则认定此处存在衍射峰。这里 T_0 取值为 1.5。

$$\begin{cases} T(i) = y'_i / \left[\sum_{i=0}^{n-1} (y'_i - \bar{y}')^2 / n \right]^{1/2} \\ \bar{y}' = \frac{1}{n} \sum_{i=0}^n y'_i \end{cases} \quad (4-4)$$

4.2.4 全谱拟合技术与数字分峰

XRD 图的分峰处理，本质上是对 XRD 图谱上的离散数据点进行数据拟合，但又和普通的数据拟合有所不同。在这里用于拟合的目标函数的类型是已知的，不是纯粹的数学上的拟合，有一定的物理意义。这里采用 Voigt 函数作为衍射峰型函数。

分峰技术的基本原理是：在数字拟合计算过程中，每一个峰的 Voigt 函数的强度、峰位、半高宽和峰型参数都作为可优化参数进行全谱拟合计算，采用 Newton-Raphson 算法以目标函数 Y 进行优化计算，以减少计算过程的误差，并以相关因子 $r^2=Y$ 作为拟合过程的终止评判指标。

$$Y = \sum_{i=1}^n (I_i - \sum_{j=1}^m I_{calik})^2 \quad (4-5)$$

(1) 式中 n 为试验点数，m 为峰数， I_i 为第 i 步测得的强度， I_{calik} 为第 k 峰第 i 步计算强度，其表达式为：

$$I_{calik} = \eta L + (1 - \eta)G \quad (4-6)$$

(2) 式中 η 为峰型可调参数，L 为洛伦兹函数，G 为高斯函数，二者分别由 (4-7) 和 (4-8) 式给出。

$$G = \frac{\sqrt{C_0}}{(H_k \sqrt{\pi})} \exp[-C_0(2\theta_i - 2\theta_k)^2 / H_k^2] \quad (4-7)$$

$$L = \frac{\sqrt{C_1}}{\pi H_k} \left(\frac{H_k^2}{H_k^2 + C_1(2\theta_i - 2\theta_k)^2} \right) \quad (4-8)$$

G 和 L 两个函数分别包含每个峰的强度、峰位和半高宽，并作为可调参数，式中 H_k 为第 k 个 Bragg 反射的半高全宽 (FWHM)。

4.3 程序开发与应用实例

4.3.1 程序开发

根据本章上述所建立的新的用于解决植物类弥散 XRD 图谱解析问题的图谱分峰处理模式，见图 4-3，以及各模块关键数学理论，基于 MATLAB 软件平台，开发 XRD 图谱数字分峰系统。该系统包括格式转化、平滑滤波、背景扣除、初步寻峰和分峰拟合等五个功能模块，详细见图 4-5。

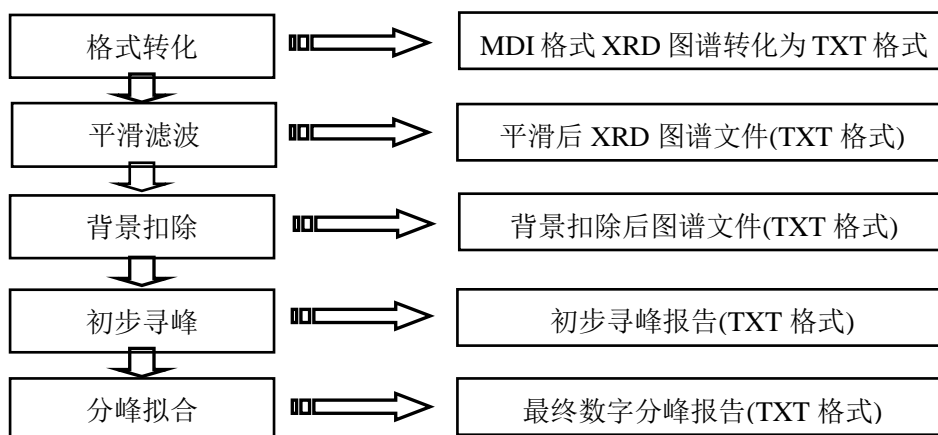


图 4-5 图谱分峰软件功能模块

软件中的五个模块既可按照图 4-5 中的流程顺序获得最终的分峰报告也可独立使用。为方便软件各个模块间的数据交流，本软件设置格式转化模块，可将 MDI 格式文件转化为 TXT 文件。这样产生的文件亦可用于其他图谱分析软件，提高了软件的适用范围。与传统 XRD 分析模式相比，本软件在处理流程上最大的不同是最后两个功能模块，即初步寻峰和分峰拟合。这也是本软件创新所在。初步寻峰模块基于对称零面积卷积理论，对图谱进行寻峰处理，获得精确的峰位信息和初步估计的峰型参数，输出初步的寻峰报告。该文件是进行分峰拟合的关键。因为分峰拟合模块是采用最小二乘法，以各独立衍射峰函数的叠加函数作为目标函数，进行曲线拟合分峰运算，而其是否成功和处理速度快慢多取决于参数初值的选择。所以初步的寻峰报告是进行下一步拟合分峰的基础，尤为重要。该软件系统初始界面见图 4-6。

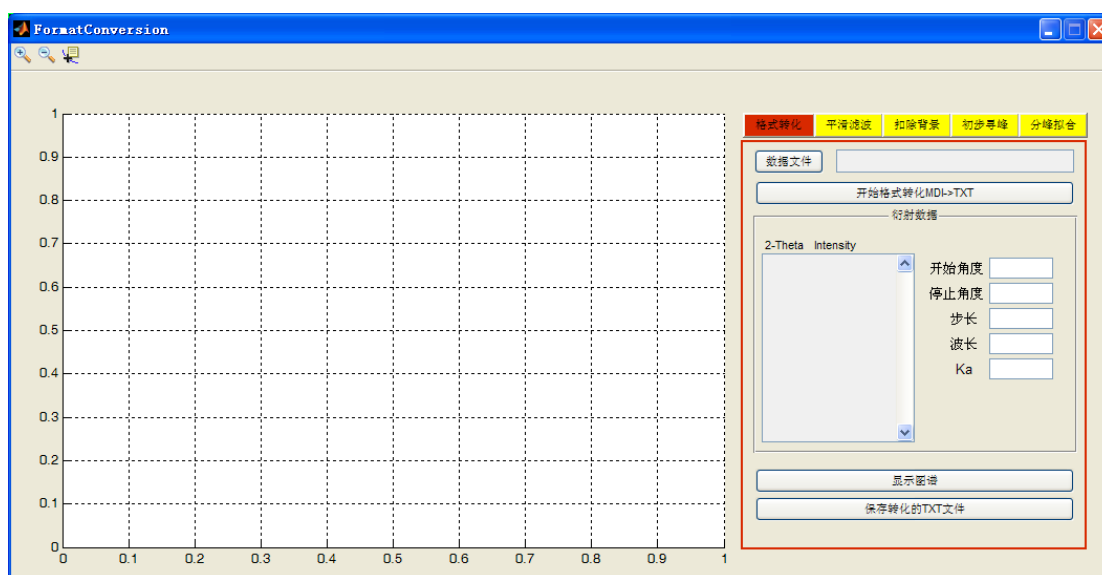


图 4-6 XRD 图谱数字分峰软件初始界面

4.3.2 应用实例

下面应用 XRD 图谱数字分峰软件对三种植物类藏药图谱进行分峰处理，药材信息见表 4-1，其 XRD 图谱见图 4-7。

表 4-1 三种藏药信息

编号	中文名称	拉丁文	采集地区
2011012g	党参	Codonopsis pilosula (Franch.) Nannf.	甘肃岷县
2011032y	菥蓂子	Thlaspi avrense L.	迪庆
201117sd	蔷薇花	Rosa multiflora Thunb	四川迭部

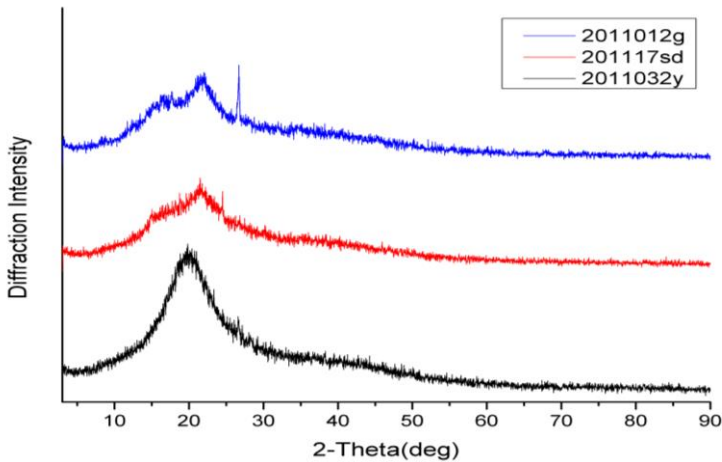


图 4-7 三种植物类藏药的原始 XRD 图谱

从图 4-7 可以看出，三种植物类藏药的 XRD 图谱均表现为重叠的弥散性衍射峰，主极大衍射峰分布于掠射角为 10~30° 的范围内，这是因为植物类藏药所含的蔗糖、蛋白质和淀粉等大部分有机物质多属于低级晶系，结晶度很低，这也导致其图谱有很多“毛刺”，和图谱噪声混杂在一起。所含微量的无机物质产生的尖锐衍射峰叠加在一个“馒头峰”之上。这些特点导致其缺乏定量化数字信息，无法对藏药进行精确唯一表征。表 4-2 是依据传统 XRD 处理模式，采用 JADE 6 图谱分析软件，对三种藏药进行寻峰处理的结果。

表 4-3 JADE6 软件寻峰结果

样品编号	d/(I/I ₀)	
2011012g	3.3357/1	
2011032y	3.3399/1	
201117sd	4.1339/1	3.6293/0.945

注：d 为晶面间距；I/I₀ 为衍射相对强度。

从表 4-3 可知，JADE 软件对药材党参和菥蓂子只能寻到一个峰，对蔷薇花也只能寻到两个峰。原始 XRD 图谱虽然包含着大量的丰富的药材特征信息，但传统

的 XRD 分析模式只能获得极为有限的信息。这时便需要借助像图谱分峰拟合等技术来对图谱进行处理,以挖掘足够特征信息。下面使用开发的 XRD 图谱分峰软件对这三种藏药进行分峰拟合,处理后的图谱见图 4-8~10。图中外轮廓中的黑色虚线代表实验点 I_i ; 外轮廓中的红色实线代表各衍射峰叠加的模拟曲线,即 $\sum I_{calik}$; 实验曲线下各峰为分峰图谱,即每个峰代表 I_{calik} 。可见数字拟合分峰从整体轮廓上与实验所得曲线较好地重合。

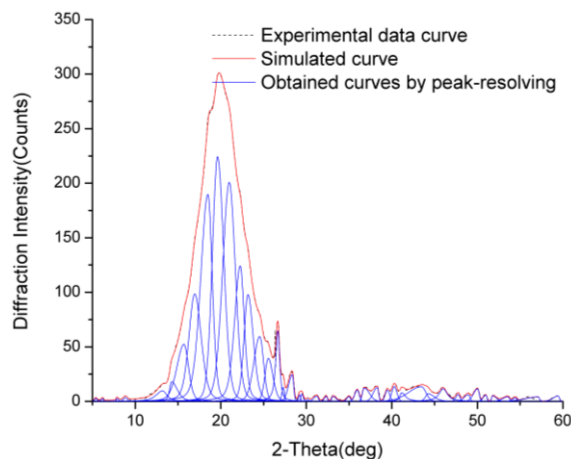


图 4-8 藏药党参的分峰图谱

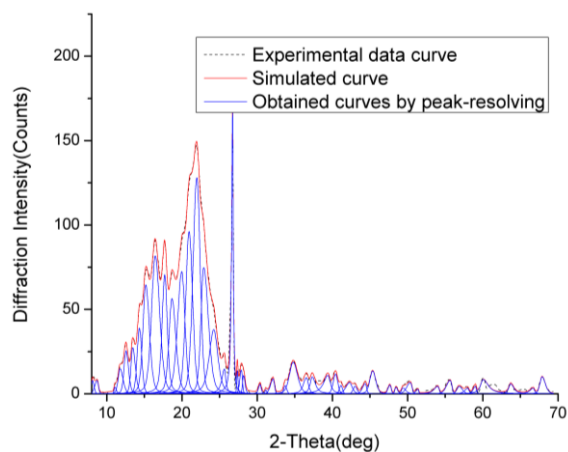


图 4-9 藏药芥冥子的分峰图谱

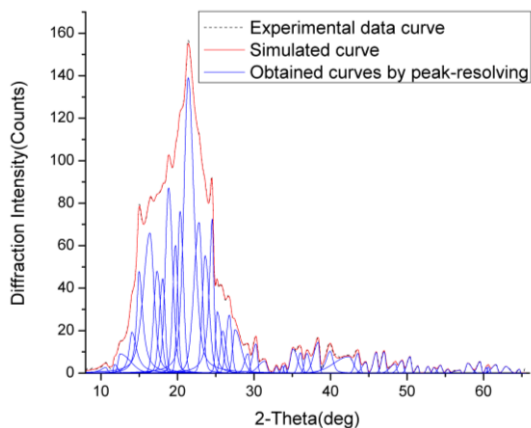


图 4-10 藏药蔷薇花的分峰图谱

经图谱分峰软件处理后,会生成一个和样品相对应的分峰报告,其中包括各衍射峰位、绝对和相对衍射峰强,半高宽、峰型因子以及偏态因子等非常丰富的谱峰信息。利用这些可以进行求取结晶度,鉴别同类药材等问题。以芥冥子为例给出其 $10\sim 30^\circ$ 范围内的衍射峰数据,见表 4-4。

表 4-4 菥冥子图谱分峰数据

峰号	峰位	相对峰强	半高宽	峰型因子	偏态因子
1	10.66	0.03252	0.6366	0.99905	0.84526
2	11.77	0.04786	0.3947	0.00269	-0.82951
3	12.51	0.09715	0.87408	0.04819	-0.92749
4	14.05	0.24012	1.02554	0.9993	-0.4954
5	14.98	0.50198	0.86521	0.99669	-0.40384
6	16.4	0.52806	1.52479	0.44775	0.44555
7	17.34	0.53912	1.08261	0.04935	-0.08149
8	18.07	0.56614	0.8482	0.11162	0.3117
9	18.85	0.65655	1.10809	0.19602	-0.15688
10	19.76	0.68645	0.76923	0.15584	0.22931
11	20.35	0.78497	0.84481	0.03656	-0.23159
12	21.4	1	1.49325	0.38911	-0.26693
13	22.82	0.71945	1.15617	0.18376	0.44509
14	23.62	0.57762	1.03084	0.0894	-0.28078
15	24.55	0.57658	0.59203	0.99262	0.55534
16	25.19	0.2812	0.64509	0.21131	-0.7401
17	25.9	0.2564	0.83215	0.0028	-0.37398
18	26.78	0.22624	0.81557	0.07626	0.241
19	27.52	0.1614	0.96315	0.12022	-0.72981
20	29.21	0.08548	0.6059	0.10466	0.65064

第五章 藏药 XRD 二阶导数指纹图相似度计算

随着现代藏药越来越广泛为人们所认识和接受,市场对中药的需求量也愈来愈大,对藏药资源的开发和保护已提上议事日程,即在注重保护野生动植物生态平衡的同时又要满足社会对藏药资源的需求。但是在藏药材资源开发研究中,目前存在一个被长期困扰的问题,即如何鉴别药材的真伪,野生还是栽培,以及如何评价药材内在品质。目前对野生和栽培藏药材的鉴别,仅限于形体、颜色、质地等外在形貌上的主观判断或从微量元素方面探讨其内在质量上的相关性。藏药指纹图谱能从整体上综合分析野生和栽培藏药材在内在质量及所含成分上的异同及相关性,从而筛选出适合临床要求的要用替代品,处理好藏药资源保护和开发之间的关系。大多数植物类和果实类藏药因为含有大量的像淀粉、蛋白质和蔗糖等之类的大分子有机成分,致使其 XRD 指纹图谱特征均呈模糊弥散性峰。因此,传统的 XRD 分析方法面临严重考验,其 XRD 指纹图谱的鉴别不能仅凭主观方法判断,而必须借助有些数学方法来描述图谱间的相似程度,揭示图谱间的微小差异。对于矿物类和化石类藏药,由于具备很好的结晶度,图谱表现出丰富而明显的特征信息,可直接用以鉴别药材真伪及是否野生,但却不能评价药材品质的差异,也需引入如相似度等数学方法。

相似度作为数字信号科学中的一个定量定性参数,已被国家药典委员会确定为中药指纹图谱标准中的一个重要评价指标,目前应用于中药注射剂及色谱类指纹图谱的研究。相似度计算一般采用两种方法,即相关系数法和向量夹角余弦法^[29,30]。利用相关系数可确定两种属性之间的关系,它强调了数值涨落的比较。为了更好地描述图谱间相似程度,人们基于这两类方法,改进目前的相似度算法,提出新的处理模式,如刘云飞等人^[31]建立的全谱数据库计算指纹图谱的相似度,王康等人^[32]建立的基于相对熵的相似度算法等,获得很好的效果。本文将对相似度理论及延伸出来的一些算法加以介绍和选择应用。

5.1 相似度理论及算法

5.1.1 夹角余弦法

所谓夹角余弦法,是从两个向量间夹角的余弦计算公式得来的。假定 n 维向量 \vec{a} 和 \vec{b} , 那么这两个向量间的夹角余弦就是:

$$\cos\langle\vec{a},\vec{b}\rangle=\frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|}=\frac{\sum_{i=1}^na_ib_i}{\sqrt{\sum_{i=1}^na_i^2\cdot\sum_{i=1}^nb_i^2}}$$

该夹角余弦范围是 $[-1,1]$, 当其值为 1 时, 说明这两个向量完全相似; 当靠近 1 时, 则两个向量非常相似; 当其值为 0 时, 则两个向量完全不一样, 相似度为 0; 而当靠近 0 时, 那么这两个向量差别很大, 也就是相似度很小。在本章节, 应用此法来计算 XRD 图谱间的相似度, 即是分别提取两个图谱的特征向量参与计算。

5.1.2 相关系数法

通常所说相关系数, 一般指变量间的相关系数, 作为刻画样品间的相似关系也可类似给出定义, 即第 i 个样品与第 j 个样品之间的相关系数定义为:

$$r_{ij}=\frac{\sum_{a=1}^p(x_{ia}-\bar{x}_i)(x_{ja}-\bar{x}_j)}{\sqrt{\sum_{a=1}^p(x_{ia}-\bar{x}_i)^2\cdot\sum_{a=1}^p(x_{ja}-\bar{x}_j)^2}} \quad -1\leq r_{ij}\leq 1$$

$$\text{其中, } \bar{x}_i=\frac{1}{p}\sum_{a=1}^px_{ia}, \quad \bar{x}_j=\frac{1}{p}\sum_{a=1}^px_{ja}$$

5.1.3 相对熵方法

相对熵又称为 KL 散度, 信息散度, 信息增益, 它是两个概率分布 P_2 和 P_1 差异性的非对称性的度量, 记为 $D(P_2, P_1)$ 定义如下:

$$D(P_2, P_1)=\sum_{i=1}^kP_2(a_i)\log\frac{P_2(a_k)}{P_1(a_k)}$$

其运算结果的单位为比特。上述定义中约定, 当 $P_2(a_k)=0$ 或 $P_1(a_k)=0$ 时,

$$\log\frac{P_2(a_k)}{P_1(a_k)}=0.$$

相对熵的一些特点罗列如下:

- (1) $D(P_2, P_1)$ 有其方向特性, 一般 $D(P_2, P_1) \neq D(P_1, P_2)$.
- (2) $D(P_2, P_1) \geq 0$

基于以上特性，两个概率分布间的散度可以定义为 $J(P_2, P_1) = D(P_2, P_1) + D(P_1, P_2)$ ，散度是两个概率分布间差异的度量，有如下三个特点：

- (1) $J(P_2, P_1)$ 没有方向特性，参与运算的两概率分布是对称的，即 $J(P_1, P_2) = J(P_2, P_1)$ 。
- (2) $J(P_2, P_1) \geq 0$
- (3) $J(P_2, P_1) = 0 \Leftrightarrow P_2 = P_1$

本文中，散度用以计算两 XRD 指纹图谱间的相似度，其结果是标准化的，详细的算法流程介绍如下。

- (1) 采集 XRD 指纹图谱数据并使其标准化；
- (2) 载入标准化 XRD 图谱数据，找到其最大值和最小值，记为 max 和 min。
- (3) 将图谱数据分割为 N 个部分，步长定义为 $h = \frac{(\max - \min)}{N - 1}$ ，获得的 N 个数据区如下， $[\min, \min+h], [\min+h, \min+2h], \dots, [\min+(N+1)h, \max]$ 。
- (4) 将 n 个加载的数据映射到 N 个部分中的 N+1 个端点数据。 $f: x_i \rightarrow y_j$ ，这里的 x_i 指被加载的数据集，而 y_j 是 N 个部分中端点数据。
- (5) 计算被加载数据在 N+1 个端点中的概率分布。
- (6) 以散度公式计算相对熵：

$$D(P_2, P_1) = \sum_{i=1}^N P_2(a_i) \log \frac{P_2(a_i)}{P_1(a_i)} + \sum_{i=1}^N P_1(a_i) \log \frac{P_1(a_i)}{P_2(a_i)}$$

如果有很多组数据，散度可以分别计算获得。

5.2 相似度计算模式及程序设计

基于上述关于相似度计算的理论，提出用以计算不同 XRD 指纹图谱间相似度的处理模式，见图 5-1。

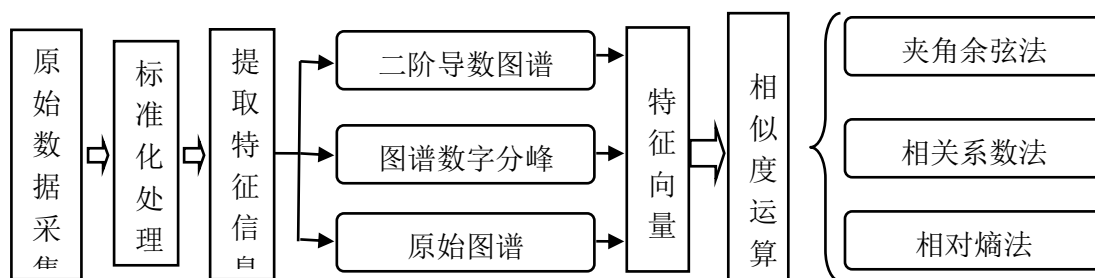


图 5-1 相似度处理模式

(1)原始数据采集：应用 X 射线衍射仪采集相应图谱，并基于 MATLAB 编写格式转化程序，将 MDI 格式的图谱文件转化为 TXT 格式文件，便于后期处理；

(2)标准化处理：这部分意在获得标准化的 XRD 图谱数据。根据图谱文件本身特点，规定用以计算相似度的衍射角范围，对 X 射线衍射强度进行归一化处理。

(3)提取特征信息：这部分有三个可供选择的处理方法用以获得表征相应藏药材的特征向量。一是，利用离散数学中的二阶导数理论对原始图谱进行信息挖掘，获得足够的指纹特征；二是，引入 Voigt 衍射峰函数模型和全谱拟合技术，对原始图谱进行分峰拟合，获得精确独立的衍射峰信息；三是直接应用原始图谱，结合相对熵算法，获得图谱间相似度。

(4)相似度运算：这里三个相似度算法可供选择，即夹角余弦法、相关系数法和相对熵算法。与前两个不同的是，相对熵法是对整个图谱进行计算，而非仅对有限的特征衍射峰。

本文使用 MATLAB 程序设计语言来编写相似度计算程序。

5.3 相似度应用实例

基于上述相似度计算理论，本章节分别建立以贝壳类为例的矿物类藏药和以大黄、蔷薇花为例的植物类藏药的标准 X 射线衍射对照指纹图谱，并对同类或不同产地药材进行相似度分析。

5.3.1 样品与实验

实验用藏药样品的编号、名称、拉丁文等信息见表 5-1。

表 5-1 实验用样品信息

	编号	中文名称	拉丁文	产地
贝壳类	2011017sc(1#)	海蛤壳	Concha Meretricis seu Cyclinae	广西
	2011010sc(2#)	珍珠母 1	Concha Margaritifera	江苏
	2011032sc(3#)	珍珠母 2	Concha Margaritifera	江苏
	2011041sc(4#)	瓦楞子	Ark Shell Concha Arcae	广西
	2011035sa(5#)	珍珠	Pernulo	四川阿坝
test	2011023sc(6#)	蜗牛	Fruticicolidae	四川
	2011050sc(7#)	鱼脑石	Pseudosciaena crocea	成都
大黄	2011008g(8#)	大黄	Rheum palmatum L.	甘肃岷县
	2011058gh(9#)	大黄	Rheum palmatum L.	甘肃合作
	2011072sa(10#)	大黄	Rheum palmatum L.	四川阿坝
	201105sd(11#)	水大黄	Rheum alexandrae Batal.	四川迭部
	2010036(12#)	大黄	Rheum palmatum L.	青海果洛
test	2011001g(13#)	麻黄草	Herba Ephedrae	甘肃岷县

将表 5-1 中的矿物类和植物类藏药样品经粉碎机粉碎，过 200 目筛，制成细粉，以备 X 射线衍射实验使用。实验所用仪器是荷兰飞利浦公司生产的 Y-4Q 型全自动 X 射线衍射仪，Jade6.5 数据处理软件。XRD 测试条件：管压 30kV，管流 20mA 铜靶的 $K\alpha$ 辐射，滤波片为镍， $DS=1^\circ$ ， $SS=1^\circ$ ，扫描速度为 0.3 度/秒，连续扫描，时间常数 0.5 秒，扫描范围 $10\sim 90^\circ$ ，步长 0.03。

5.3.2 贝壳类矿石藏药

下面详细介绍建立贝壳类藏药标准 XRD 对照指纹图谱，并进行相似度分析的过程。

(1) 首先观察和分析实验获得的贝壳类样品的 XRD 图谱，并进行归一化处理，并选取 $10\sim 90^\circ$ 范围作为研究区域，见图 5-2。可以看出贝壳类图谱含有很多尖锐的衍射峰，峰位明确，显示出很强的图谱特性。1#~5#图谱具备一致的几何拓扑规律，但在衍射角方向上有微小位移偏差，6#和 7#虽然和其轮廓大致相同但部分峰位不一致。

(2) 然后应用 XRD 图谱分析软件对样品进行寻峰处理。寻峰时，统一平滑和扣除背景等参数，以保证数据可比性。实验数据以晶面间距 $d(A)$ 与相对衍射强度 I/I_0 表示，记为 $d/(I/I_0)$ ；相应的衍射峰值如下：

1#: 3.3904/83.2, 3.2662/46.8, 2.8689/17.4, 2.6973/100, 2.4819/46.0, 2.4045/17.6, 2.3716/49.3, 2.3290/34.7, 2.1864/13.2, 2.1020/18.7, 1.9751/46, 1.8788/28.4, 1.8135/22.6, 1.7430/40.2, 1.7257/23.1, 1.4107/9.6, 1.3592/10.7.

2#: 3.3841/36.9, 3.2627/21.2, 2.8656/31.9, 2.6965/100, 2.4783/27.8, 2.4032/8.6, 2.3684/29.0, 2.3258/16.5, 2.1874/5.6, 2.1030/7.9, 1.9748/17.4, 1.8759/14.8, 1.8127/11.6, 1.7413/33.1, 1.7239/17.8, 1.4124/12.1, 1.3586/8.1, 1.3504/5.9.

3#:3.3834/38.7, 3.2629/21.2, 2.8646/32.4, 2.6958/100.0, 2.4790/27.6, 2.4042/4.2, 2.3684/31.7, 2.3251/16.1, 2.1870/5.1, 2.1011/5.9, 1.9744/21.6, 1.8763/14.1, 1.8114/10.4, 1.7421/32.6, 1.7239/19.4, 1.4132/10.0, 1.3592/6.8, 1.2239/4.0, 1.1083/4.7

4#: 3.3871/100, 3.2659/28.9, 2.8639/18.7, 2.6950/83.5, 2.4817/39.2, 2.3700/60.3, 2.3274/27.5, 2.1865/11.7, 2.1008/21.3, 1.9749/54.5, 1.8779/25.1, 1.8137/18.4, 1.7414/41.1, 1.7249/21.3, 1.4135/9.3, 1.3594/8.1, 1.2612/8.1

5#: 3.3834/40.4, 3.2652/22.8, 2.8634/26.9, 2.6944/100, 2.4798/29.7, 2.3681/33.0, 2.3256/16.2, 2.1861/6.0, 2.1005/8.3, 1.9728/20, 1.8748/15, 1.8126/15, 1.7412/34.4, 1.7245/19, 1.4120/11, 1.3583/8.4, 1.2604/4.1, 1.2236/5.0, 1.2058/4.3, 1.1039/4.1

6#: 4.2477/21.9, 3.3806/76.1, 3.3319/86.9, 3.2633/33.3, 3.1826/50.3, 3.0251/56.2, 2.8633/24.5, 2.6964/65.7, 2.4759/100, 2.3650/37.9, 2.3229/21.9, 2.2794/19.6, 2.1843/19.9, 2.1016/23.5, 1.9738/46.7, 1.9070/14.1, 1.8734/38.6, 1.8128/20.9, 1.7394/34.6, 1.7232/19.6, 1.5564/15.4, 1.4113/12.7, 1.3804/10.5, 1.2390/23.5

7#: 3.3355/73.4, 3.2106/46, 2.6646/40.3, 2.4489/82.3, 2.3470/33.1, 2.3156/97.6, 2.1667/29.8, 2.0876/85.5, 1.9605/100, 1.8630/54.0, 1.7990/26.6, 1.7148/26.6, 1.5475/21.8, 1.4551/20.2, 1.3939/20.2, 1.2342/30.6, 1.1997/18.5, 1.1657/32.3

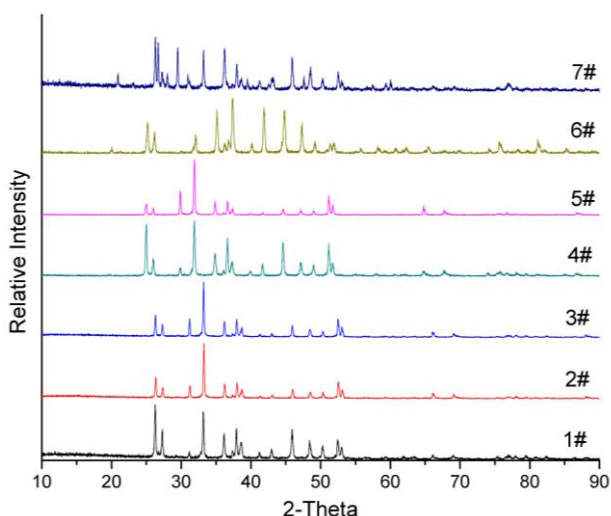


图 5-2 矿石类示例藏药图谱

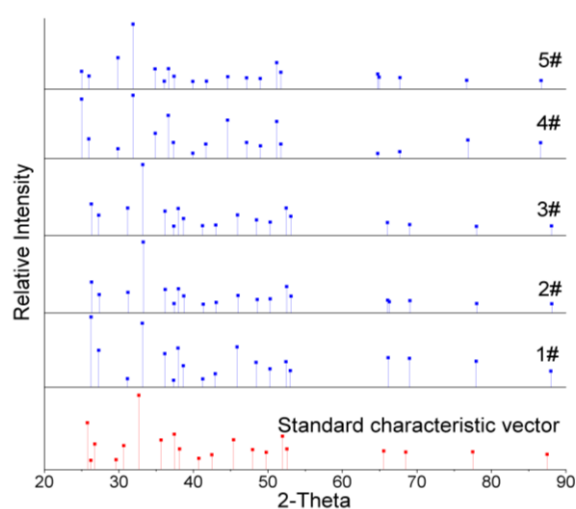


图 5-3 对照和各图谱衍射峰特征向量图

(3) 接着就是提取图谱特征信息, 建立贝壳类藏药的标准 XRD 对照指纹图谱。这里沿用“共有峰模式”来计算图谱相似度, 即根据已知多个 X 射线衍射图谱的几何拓扑规律进行峰的匹配, 无对应峰的以零补充, 然后对衍射峰某一特征属性进行某种数学处理如均值或中位值, 获得相应的平均值或中位值图谱, 以此作为比照来计算不同药材图谱与对照图谱间的相似度。我们采用平均值处理方式, 对 1#~5#等 5 个几何图谱规律非常一致的 XRD 图谱, 获得贝壳类标准 XRD 对照指纹图谱, 见图 5-3。对照指纹图谱中, 采集其峰位属性作为对照特征向量, 其它已知图谱亦是如此, 见表 5-2。

(4) 最后根据所得各图谱峰位特征向量, 利用夹角余弦法和相关系数法, 计算相应的相似度, 结果见表 5-2。

5-2 贝壳类藏药 XRD 图谱相似度计算结果

编号	1#	2#	3#	4#	5#	6#	7#	对照图谱
1	26.274	26.336	26.319	25.014	25.001	26.304	25.194	25.7888
2	27.323	27.356	27.31	25.961	25.992	27.282	26.156	26.7884
3	31.177	31.241	31.197	29.874	29.904	30.957	0	30.6786
4	33.147	33.264	33.206	31.899	31.914	33.219	32.108	32.686
5	36.159	36.221	36.205	34.899	34.869	36.221	35.111	35.6706
6	37.347	37.407	37.372	0	36.095	36.664	36.263	29.6442
7	37.93	38.004	37.96	36.652	36.655	37.989	36.836	37.4402
8	38.65	38.723	38.695	37.315	37.401	38.663	37.39	38.1568
9	41.26	41.337	41.245	39.924	39.94	41.261	40.15	40.7412
10	42.956	43.059	43.013	41.666	41.768	43.03	41.89	42.4924
11	45.879	45.985	45.925	44.591	44.634	45.91	44.859	45.4028
12	48.416	48.579	48.475	47.155	47.156	48.564	47.335	47.9562
13	50.292	50.322	50.332	49.014	48.985	50.318	49.222	49.789
14	52.436	52.526	52.481	51.189	51.189	52.511	51.4	51.9642
15	53.033	53.125	53.079	51.744	51.759	53.091	51.94	52.548
16	66.143	66.073	66.056	64.748	64.752	66.066	65.499	65.5544
17	0	66.295	0	0	64.943	0	0	26.2476
18	69.021	69.085	69.043	67.691	67.749	0	0	68.5178
19	77.962	78.014	78.008	76.841	76.702	76.902	0	77.5054
20	87.974	88.091	88.057	86.595	86.684	83.846	0	87.4802
夹角余弦	0.9922	0.9847	0.9922	0.9833	0.9845	0.9409	0.7605	1
相关系数	0.9500	0.8684	0.9499	0.9330	0.8683	0.6455	0.1413	1

(5) 分析: 从表 5-2 中的相似度值可以看出, 样品 1#~5#与建立的贝壳类标准对照图谱相比, 具有很高的相似程度, 都可以归入到贝壳类藏药。物相检索获得的结果也能佐证这一结论, 样品 1#~5#主相皆为碳酸钙, 只是由于产地及环境的不同导致了微小差异。观察参与测试和比较两个药材 6#和 7#, 可以看出

6#具有较高的夹角余弦值和不低相关系数值,可以考虑将其归入贝壳类,这也与物相检索所得结果一致,但 7#样品却不同,其夹角余弦值为 0.7605,较低,加上其相关系数更低,仅为 0.1413,不能归入贝壳类。物相检索证明其主相为 Hafnium Manganese Silicon(HfMnSi),自然与贝壳有较大差异。通过以上分析可以看出,夹角余弦和相关系数可以用来度量图谱间相似性。

5.3.3 不同产地大黄图谱相似度分析

下面详细介绍建立藏药大黄的标准 XRD 对照指纹图谱,利用 KL 散度,即相对熵理论,对其相似度分析的过程。

(1) 观察分析:图 5-4 显示了 6 个参与计算的藏药样品 XRD 图谱。可以看出,图谱都含有处在 $10\sim 30^\circ$ 范围内的弥散“馒头峰”,并重叠有较为尖锐的衍射峰,8#~13#尖锐衍射峰部分相似,但前 5 个总体几何拓扑规律非常一致,13#的馒头峰部分较其他差别较大。选择 $3\sim 70^\circ$ 衍射角范围,作相似度计算。

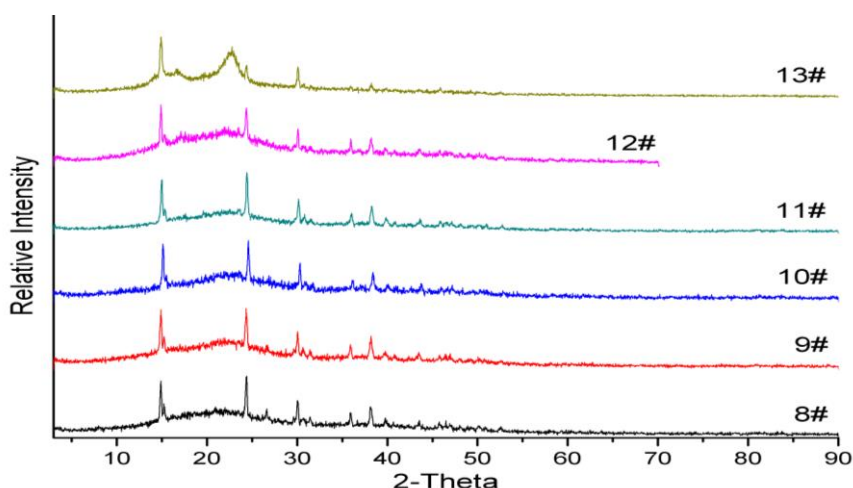


图 5-4 藏药大黄及测试用样品 XRD 图谱

(2) 图谱预处理:首先对各图谱进行归一化和平滑去噪处理,这里采用小波多区间阈值去噪算法。接着便是建立大黄的标准 XRD 对照指纹图谱,分两种情况。第一,当以衍射强度作为特征属性进行相似度计算时,建立标准图谱前要对各图谱峰位进行左右平移校正,以使各图谱间欧氏距离最小,获得客观的相似度值。第二,当以衍射角度作为特征属性时,不必矫正,按原始图谱衍射角度,建立标准图谱。见图 5-5 和图 5-6。

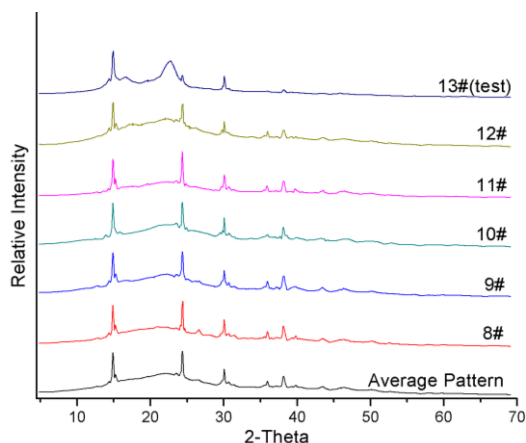


图 5-5 预处理后的样品图谱(强度属性)

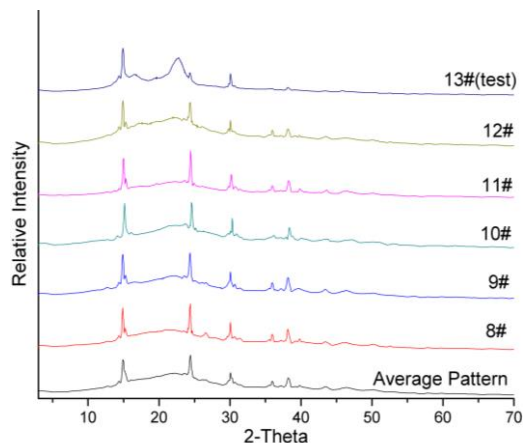


图 5-6 预处理后的样品图谱(角度属性)

(3) 计算概率分布曲线：这一步骤是后面计算相对熵的基础。利用概率统计相关知识，在(2)步骤基础上，分别获得藏药大黄图谱及其平均模式和测试样品图谱的概率分布曲线。对于特征属性是强度时，根据核平滑密度估计理论，估计各图谱样本概率密度分布，MATLAB 内置的 `ksdensity` 函数可以解决；而特征属性时衍射角度时，所要求的概率分布是衍射峰在角度上概率分布，这里考虑将强度数据转化为概率，获得图谱的角度概率分布。结果见图 5-7 和 5-8。

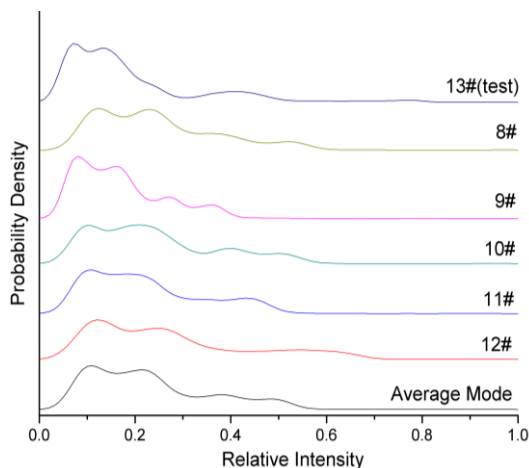


图 5-7 样品图谱的强度概率密度分布

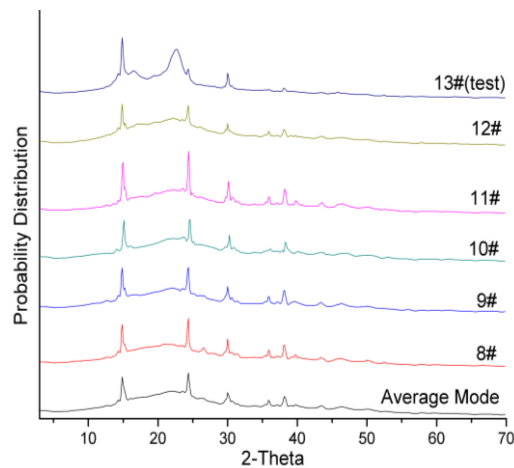


图 5-8 样品图谱的角度概率密度分布

(4) 相对熵和 KL 散度计算：以概率分布曲线 `Average Mode` 为标准，分别计算各图谱对应的概率分布的相对熵及 KL 散度值，结果见表 5-3，其关键 MATLAB 代码如下：

```
function KLDiv = KLDiv_Calc(p1,p2)
```

```
%% =====%
% 该程序用于计算两概率分布间的KL散度，以度量其差异性。
% 输入：两个概率密度分布向量p1,p2
% 输出：KL散度值KLDiv
%% =====%
REntropy1 = sum(p1.* (log2(p1./p2)));% 概率分布相对熵R(P1,P2)
REntropy2 = sum(p2.* (log2(p2./p1)));% 概率分布相对熵R(P2,P1)
KLDiv = REntropy1 + REntropy2;
end
```

表 5-3 相对熵理论求得的图谱 KL 散度值

特征属性	8#	9#	10#	11#	12#	13#(test)	对照图谱
衍射强度	0.0578	0.0286	0.0975	0.6752	0.4187	0.5932	0
衍射角度	0.00568	0.00694	0.02138	0.01159	0.00986	0.10678	0

(5) 相似度分析：这里采用相对熵理论来计算图谱间相似程度，实质上是通过图谱对应的概率密度分布的差异性来体现的，即 KL 散度值的不同。KL 散度值域位于 0 至无穷大之间，其值越小，两图谱间相似度就越小，当为 0 时，则说明图谱完全一致。考量图谱衍射强度特征属性时，其 KL 散度值于 0.0286~0.6752 间，故当某一藏药图谱位于这一范围内，或小于 1，都可认定为大黄药材。而本文所选测试样品恰巧如此。但所选药材麻黄草为麻黄科、麻黄属，与大黄(蓼科、大黄属)是不同的药材，故以衍射强度作为特征属性所得的 KL 散度值不可靠。而以衍射峰角度作为特征属性计算获得的 KL 散度值能够体现出这一事实，见表 5-3。观察大黄与麻黄草原始图谱可以看出其尖锐衍射峰很相似，所不同的是大分子有机物质造成的馒头峰部分差异较大，以衍射角度为特征属性所得相似度值正表现了这点。因为相对熵理论是整体上对图谱处理，以度量图谱间整体上的相似程度，而不再像夹角余弦和相关系数法那样对专有特征衍射峰进行计算。

第六章 总结与展望

6.1 成果与收获

本文为了解决藏药材尤其是植物类藏药 X 射线衍射图谱难以解析的问题,以传统的图谱分析模式为基础,研究和设计了三种数据处理模式,即藏药粉末 X 射线衍射衍射文件获取流程,有效实现了对藏药指纹特征信息的提取,为进一步开发相关数据库用以鉴定药材和评价质量奠定了数据基础;藏药材 XRD 全谱分峰系统,能方便快捷高效的对 XRD 图谱进行数字化分峰,挖掘丰富的图谱特征信息,为相似度和结晶度计算提供了依据;藏药 XRD 图谱相似度计算模式,能准确计算不同产地不同种类的藏药材 XRD 图谱进行相似度评价,进而对其分类和甄别。整个研究过程涉及的学科知识较广,既有物理学 X 射线衍射实验等材料科学方面的知识,也有数值微分等离散数学方面的知识,还有计算机编程。通过整个研究过程,我对 X 射线衍射图谱数据处理方面的知识有了更为深刻的认识,也了解到各个学科间的交叉之处具备很大的发展潜力。科学研究不仅需要深厚的专业素养,敏锐的观察问题的能力,还需要毅力以及坚定的信心。

6.2 展望

本文所提出的三个 XRD 图谱数据处理模式,经实例验证,达到了当初的目标,即有效解析藏药材 XRD 图谱,但是仍然存在着一些不完善的地方,需要以后进一步的改进。

1、各处理流程中的图谱平滑模块设计到的算法并没有实现自适应的对 X 射线衍射图谱进行平滑处理,更多的需要加入人工操作。而且面对不同类型的 XRD 图谱,其处理效果有差异;以后可以考虑将多个去噪算法集成起来,做到有选择的进行平滑处理。

2、本文大部分的程序都是应用 MATLAB 语言编写,以脚本文件实现的,执行起来较为繁琐;还有利用 MATLAB 的 GUI 模块编写的图谱分峰系统,虽然具备交互式的功能,但是因 MATLAB 在图形界面设计能力方面较差,该系统运行缓慢,稳定性不足。以后可以尝试应用 Python 或者 C++语言对各个处理模式进行 GUI 编程,使得界面更友好,速度更舒畅,稳定性更高。

参考文献

- [1] 杨红霞, 马芳, 杜玉枝, 等. 藏药川西獐牙菜及其不同提取物的红外光谱分析[J]. 光谱学与光谱分析. 2014.
- [2] 刘震东, 宋淼, 达番琼, 等. 藏药二十味沉香丸的质量标准研究[J]. 华西药学杂志. 2003, 18: 462-464.
- [3] 吴红彦, 樊秦, 纪兰菊. 藏药石榴健胃片质量控制方法的研究[J]. 中国实验方剂学杂志. 2011, 17: 70-71.
- [4] 吕扬, 王钢力. 中药材 X—射线衍射图谱研究[J]. 药学报. 1997: 193-198.
- [5] 周俊国, 吕杨, 郑启泰, 等. 中药蛇床子的粉末 X 衍射分析[J]. 中草药. 1999: 59-61.
- [6] 王树春, 吕杨, 吴楠, 等. 中药材熊胆的 X 衍射 Fourier 谱分析[J]. 中草药. 2000, 31: 214-215.
- [7] 郑笑为, 江仁望. 中药材珍珠的 X 衍射 Fourier 谱研究[J]. 药物分析杂志. 1999: 246-251.
- [8] 李岑, 桑老, 楞本才让, 等. 藏药珠西的化学成分与结构分析[J]. 光谱学与光谱分析. 2012, 32: 1671-1673.
- [9] 赵旭东, 胡延萍, 谢小龙, 等. 唐古特大黄的粉末 X 射线衍射图谱[J]. 安徽农业科学. 2007, 35: 1705.
- [10] 全正香, 魏立新, 杜玉枝, 等. 藏药南寒水石结构成分及热稳定性分析[J]. 中国中药杂志. 2011, 36: 691-693.
- [11] Liu, Y., Duan, X.W., Hu, C.X., Sun, H. and Zhu, J.R. Comparison Research about Wavelet and Moving Smoothing Method Applied in Denoising X-Ray Diffraction Patterns of Tibetan Medicine. Applied Decisions in Area of Mechanical Engineering and Industrial Manufacturing, 2014.
- [12] 胡耀垓, 张晓星, 王震, 等. SF₆ 气体分解组分红外光谱信号的去噪与背景扣除[J]. 高电压技术. 2011, 37: 1166-1171.
- [13] 方勇, 曾立波, 雷俊锋, 等. 一种新的 X 射线能谱背景扣除方法[J]. 分析测试学报. 2001, 20: 23-27.
- [14] 龙斌, 冯天成, 苏川英, 等. 一种 γ 能谱散射本底的自适应扣除方法[J]. 核电子学与探测技术. 2013: 1293-1296.
- [15] 杨建华, 胡恩萍, 郭灵虹, 李晖. 中药 XRD 二阶导数指纹图谱的研究[J]. 天然产物研究与开发, 2006, 18(3): 390-393
- [16] 夏爱生, 陈博文, 胡宝安, 王瑞, 王强. 二阶三点数值微分公式的外推算法[J]. 天津理工大学学报, 2005, 21(6): 37-39
- [17] 王艳. 二阶导数的五点数值微分公式及外推算法[J]. 天津理工大学学报, 2009, 25(4): 37-39

- [18] 许小勇, 钟太勇. 三次样条插值函数的构造与 Matlab 实现[J]. 自动测量与控制, 2006, 25(11):76-78
- [19] 罗煦琼. 一维搜索问题的三次 B 样条插值法[J]. 浙江科技学院学报, 2008, 20(1):1-3
- [20] 张琳, 聂孟喜, 仝辉. 三次和五次 B 样条函数在动力响应分析中的应用[J]. 清华大学学报(自然科学版), 2006, 46(3):327-330
- [21] 蔡天净, 唐瀚. Savitzky-Golay 平滑滤波器的最小二乘拟合原理综述[J]. 数字通信, 2011, 38(1):63-68
- [22] 杜云朋, 王建斌, 靳小强. 超声导波管道检测的小波模极大值去噪法[J]. 电子测量与仪器学报, 2013, 27(7):683-687
- [23] 董璐璐, 房文静, 徐静. 基于小波模极大值的测井信号滤波[J]. 测井技术, 2012, 36(2):141-145
- [24] 王翔, 朱正林, 田永伟. 基于小波变换模极大值的汽轮机振动信号去噪[J]. 南京工程学院学报(自然科学版), 2010, 8(1):38-42
- [25] 谭帅, 祝忠明, 周潞. 自适应空域相关滤波法消噪在 CSEM 中的应用[J]. 中国科技信息, 2013(10):160-161
- [26] 付炜, 许山川. 一种改进的小波域阈值去噪算法[J]. 传感技术学报, 2006, 19(2):534-536, 540
- [27] 庞巨丰. γ 能谱的数据分析[M]. 西安:陕西科学技术出版社, 1990
- [28] 毕云峰, 李颖, 郑荣儿. LIBS/Raman 光谱对称零面积变换自动寻峰方法研究[J]. 光谱学与光谱分析, 2013, 33(2):438-443
- [29] 张泰铭, 赵哲, 方宣启, 等. 利用样本成分耗散物的非线性化学指纹图谱原理及相似度计算与评价 [J][J]. 中国科学: 化学. 2011, 41(10): 1604-1621.
- [30] 蔡利, 李秋潼, 黎洪利, 等. 烟用香精指纹图谱相似度评价方法的选择[J]. 安徽农业科学. 2010(23): 12925-12926.
- [31] 刘云飞, 周利, 张春风, 等. 基于全谱数据库计算指纹图谱相似度新方法的探索研究 [J]. 中医药学报. 2012, 40(2): 49-51.
- [32] 王康, 杜凯, 李华. 相对熵方法用于中药指纹图谱相似度计算 (英文)[J]. 计算机与应用化学. 2007, 1: 14.

致谢

转眼间三年的研究生学习生涯就要结束了。这三年里我学到了较之以往更多的专业知识，经历了很多从未涉足的事情，但更多的是体验到了来自于家人，老师和朋友的关怀和支持，是他们的存在让我的三年光阴纵然充满荆棘也能面带笑容淡定走过。在这里，我向他们表达最真挚的感谢，这份感谢不会随着时光流逝，它会始终不渝，鲜如昨日。

我也感谢青海师范大学，感谢这里的花花草草，一砖一瓦，她们是我人生的一部分，我的人生因此而变得丰富、深刻和立体。

我能顺利地完 成论文，主要得益于我的导师段新文教授的指导和教诲。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。段老师不仅在学业上给我以精心指导，同时还在思想、生活上给我以无微不至的关怀，在此谨向他致以诚挚和崇高的谢意。

在此，我要衷心感谢胡老师，杨老师，马老师给予我实验和专业知识上的诸多帮助和指导，也感谢各位实验室的学弟学妹们在研究生三年里给予的生活和学习上的帮助。

由于本人能力有限，本文难免存在不足，请各位老师、专家批评指正，谢谢！

在校期间的研究成果及发表的学术论文清单

发表论文：

1. Yuan Liu, Xinwen Duan, Chengxi Hu, Hao Sun and Jianrui Zhu. Comparison Research about Wavelet and Moving Smoothing Method Applied in Denoising X-ray Diffraction Patterns of Tibetan Medicine. Applied Mechanics and Materials, Vol. 577 (2014), pp. 771-776
2. Yuan Liu, Zhongshan Hu, Chengxi Hu. X-ray diffraction second derivative spectrum of Tibetan medicine based on the spline interpolation method. ICMT.2011 "Multimedia Technology"(ISBN: 978-1-61284-771-9)
3. Chengxi Hu, Yuan Liu, Peng Liu, Weiwei Zhang, Jianrui Zhu. Microwave dielectric properties of (1-x) SiO₂-xTiO₂ composite ceramics derived from core-shell structured microspheres. Materials Research Bulletin. Volume 53, May 2014, Pages 54-57
4. DUAN Xin-wen, LIU Yuan, HU Cheng-xi, LI Yong-de. The Peak Resolution Of Diffuse X-ray Diffraction Pattern Of Tibetan Medicine. Advanced Materials Research. Vols. 562-564 (2012) pp 1959-1963
5. 胡成西, 刘远, 胡忠山, 李永得, 张志良, 杨维丰. 衍射端悬挂滤波片对硅 X 射线衍射图谱的影响[J]. 硅酸盐通报, 2011, 30(1): 226-229
6. 段新文, 刘远, 孙浩. 植物类藏药 X 射线衍射图谱全谱分峰拟合软件的设计与应用[J]. 内蒙古学报(自然科学版), 2015, 46(1): 48-54

专利情况：

1. 刘鹏, 胡成西, 刘远. 近零谐振频率温度系数二氧化硅基复合陶瓷及其制备方法. 中国国家发明专利(已授权), 专利号: ZL201310283214. X
2. 刘鹏, 胡成西, 刘远. 低温烧结二氧化硅基复合陶瓷及其制备方法. 中国国家发明专利(已授权), 专利号: ZL201310283083. 5
3. 杨维丰; 刘远; 胡中山等. 一种获得植物类药材粉末衍射文件(PDF)的方法. 中国国家发明专利(实质审查), 申请号: 201410105727. 6

专著情况：

1. 胡成西, 胡忠山, 刘远. 藏药 X 射线衍射指纹图谱网络数据库[M]. 西宁: 青海人民出版社, 2014 年

参与项目情况：

1. 国家自然科学基金: 藏药 X 射线衍射指纹图谱专家系统网络的构建, (项目号: 11064011), 主要研究者, 已结题, 2010. 7-2015. 3.