Chemometric Approach in Quantification of Structural Identity/Similarity of Proteins in Biopharmaceuticals[†]

Š. Župerl,[‡] P. Pristovšek,[§] V. Menart,[||,⊥,#] V. Gaberc-Porekar,[||] and M. Novič*[,‡]

Laboratory of Chemometrics, Laboratory of Biotechnology, and Laboratory for Biosynthesis and
Biotransformation, National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia, and Lek
Pharmaceuticals d.d., Verovskova 57, 1526 Ljubljana, Slovenia

Received November 20, 2006

We present a chemometrics study in which we show the identity or degree of similarity of 3D protein structures of various G-CSF (Granulocyte Colony-Stimulating Factor) isolates. The G-CSF isolates share the same amino acid sequence, but the preparation was carried out by somehow diverse technologies. The comparison of 3D structures was made on the basis of 2D NMR NOESY (Nuclear Overhauser Enhancement Spectroscopy) spectra of proteins. In searching for the most appropriate criteria to determine the identity or degree of similarity of selected spectral regions of different isolates, two methods for quantitative evaluation of identity/similarity were used. The first method compares all peaks in the two investigated protein spectral regions; the extent of peaks that overlap is determined. The second method includes spectral invariants originating from graph theory. The criteria of identity/similarity were calculated from graphs, derived from a collection of up to 200 peaks of investigated 2D NMR spectral region. The peaks were linked into a graph according to the sequential nearest neighborhoods. According to the first method all peaks were relevant, considering that spectral noise was previously removed; the largest similarity was found between the protein of a commercially available G-CSF drug and one of the three new isolates produced in the laboratory. The second method indicated that the pairwise similarity of the three new isolates is larger than the similarity of any of the new isolates with the commercially available drug. This is an expected result taking into account that the new isolates are produced by the same technology, while the commercial product has additives for long-term storage that could not be completely compensated. The proposed measure of similarity may help the developers of biosimilar products to optimize the controllable parameters of the production technology and eventually to argue the identity of the new isolate in comparison with the originator commercial product.

1. INTRODUCTION

Chemometrics methods are suitable to be applied to different research fields with the aim of visualization, quantification, and comparison of complex data. In pharmaceutical research area various chemometrics approaches have been mainly applied for drug design.[1] The capacity of chemometrics methods can be furthermore exploited for the investigation of proteins in biopharmaceuticals. Nowadays the pharmaceutical industry has an increased interest in the field of biopharmaceuticals. A substantial part of the FDA-approved drugs belongs to this class of drugs. Biopharmaceuticals consist of (glyco)proteins, and as such their full biological activity depends on their correct shape based on secondary, tertiary, and even quaternary structures. Consequently, the formulation and handling of biopharmaceuticals need particular attention to achieve optimal therapeutic effect and minimal adverse reaction. Since the conformational structure of a protein is easily disturbed, these structures are difficult to fully define. Different critical elements in the development of biosimilar products have to be tackled, which are not necessary in the case of the low molecular weight drugs. Nevertheless, the emerging importance of the issue of biosimilar products calls for a quantitative measure to be employed for comparison of new products with their parent (commercially available) biopharmaceuticals.

In the present study we compared several isolates of G-CSF (Granulocyte Colony-Stimulating Factor), a growth factor, which stimulates the bone marrow to produce more white blood cells. Naturally the growth factors are proteins produced in the body, but they can also be made as a drug.[2−6] G-CSF can be given to stimulate the bone marrow to produce new white cells more quickly after chemotherapy. There are several commercially available drugs with G-CSF. We compared the isolates produced in our laboratory with one of the commercially available drugs. Our goal was to find a measure for the similarity of different isolates of the protein, which have the same secondary and presumably also tertiary structure. The experimental evidence for the 3D structural information of proteins was obtained by 2D NMR spectroscopy.[7−9] 2D NMR NOESY spectroscopy[7] is widely

[†] Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

* Corresponding author phone: 386 1 4760 253; fax: 386 1 4760 300; e-mail: marjana.novic@ki.si. Corresponding author address: National Institute of Chemistry, Hajdrihova 19, P.O.B. 660, SI-1001 Ljubljana, Slovenia.

[‡] Laboratory of Chemometrics, National Institute of Chemistry.

[§] Laboratory of Biotechnology, National Institute of Chemistry.

[||] Laboratory for Biosynthesis and Biotransformation, National Institute of Chemistry.

[⊥] Lek Pharmaceuticals d.d.

[#] Deceased in February 2007.

10.1021/ci6005273 CCC: $37.00 © 2007 American Chemical Society
Published on Web 04/26/2007

used for the determination of the structure of proteins with a mass of up to 10 kD. The NOESY experiment uses the dipolar interaction of spins (the nuclear Overhauser effect, NOE) for correlation of protons. The distance *r* between two protons influences the correlation, which is usually observed for distances below 5 Å. The intensity of the NOE is in first approximation proportional to $1/r^6$. With other words, in 2D NOESY spectra we observe pairwise correlations of all protons that are sufficiently close in space. It is important that we detect correlations also for the protons, which are distant in the amino acid sequence but close in space due to protein tertiary structure. For this reason 2D NOESY is one of the most useful methods for the determination of protein 3D-structures.

The fingerprints (2D NMR NOESY spectra) of different isolates were examined, and the spectral regions characteristic for proteins were defined. Characteristic spectral regions of the isolates were compared, and the similarity criteria were calculated. Two different approaches were considered for the determination of similarity/identity criterion. One is based on direct statistics of matching peaks in the relevant spectral regions, while the other is deduced from graph theory.[10] For the latter criterion the NMR peaks have to be connected into a 2D graph with nodes (peaks coordinates) and edges (connections between the peaks) determined on the intensity-ordered list of peaks by the sequential nearest neighbors method.[11]

2. EXPERIMENTAL AND METHODS

2.1. Preparation of Samples. Different isolates of G-CSF, a recombinant protein from *Escherichia coli* (*E. coli*), were used in this study. G-CSF is a monomer globular protein containing 175 amino acids, which construct four compact alpha helices. It is stable in water solutions with low ionic strength, at pH close to 4. Using a recombinant high G-CSF producing *E. coli* strain,[6] three different isolates designated as F1, F2, and F3 were prepared. For isolation and purification of these isolates, the same procedure involving mild extraction of inclusion bodies from *E. coli* and a set of three consecutive chromatographic steps was used. However, F1 was prepared at laboratory scale, while F2 and F3 were prepared at pilot scale. Additionally, the final compositions used for long-term storage were different. F1 was formulated in pure water, acidified to pH 4.4 by addition of acetic acid. In contrast, F2 and F3 were formulated in 10 mM acetic acid pH 4.0 with 5% sorbitol added. On the other hand, G1 derived from a commercially available biopharmaceutical product containing original additives for long-term storage such as 5% sorbitol and 0.004% Tween 80. Prior to NMR spectroscopy experiments, all samples containing sorbitol were dialyzed against 10 mM acetic acid pH 4.0, and the F1 isolate was diluted using the same solvent. Protein samples for NMR were prepared at concentrations of approximately 0.7 mg/mL with D_2O added to 2.5%.

2.2. 2D NMR NOESY Spectra. NOESY spectra of four G-CSF isolates prepared as described above were recorded. The NMR experiment was carried out at 800 MHz magnetic field, with 64 repetitions at 2048 × 700 points, and a mixing time of 150 ms. The spectra were apodized with a squared sine bell function shifted by $\pi/2$ in both dimensions. The spectral peaks were picked by FELIX software,[12] positive signals only, with optimization; the threshold was manually chosen well above the noise level.

2.3. Graph Invariants with the Sequential Nearest Neighbors Method. Graph theory is a mathematical branch initiated by a well-known consideration of the Problem of the Seven Bridges of Koeningsberg by Leonhard Euler in 1736.[13] In contrast to topology where metrics does not exist, but only the concept of the neighborhood, graph theory allows metrics. Graph theory was successfully applied to chemistry (molecular structure) from 1947 on, when both Platt,[14] on one side, and Wiener,[15] on the other side, have considered molecular path numbers and closely related quantities as molecular descriptors. Chemical graph theory has advanced considerably over the past 20 years from considering isolated correlations of a structure−property relationship by using ad hoc descriptors to a well structured branch of mathematical chemistry which has made an impact on current QSAR and QSPR studies.

In this study we applied the Sequential Nearest Neighbors Method[11] to evaluate parts of 2D NMR NOESY spectra. The peak coordinates are the basis for calculation of Euclidean distances between the peaks from the intensity-ordered sets. On this way individual peaks are connected into a graph from which a distance matrix is constructed. The invariants such as row sums or eigenvalues of diagonalized matrices may be used to characterize the NMR spectral region, from which the graph has been constructed. This method has been already applied in the analysis of 2D proteomic maps.[11,16−20] The detailed procedure of the applied sequential nearest neighbors method is described elsewhere.[11] Here we summarize the main steps only:

- The sets of peaks of selected spectral regions have to be extracted and ordered by decreasing intensity for each isolate separately;
- Each set of peaks was normalized to the maximum peak intensity;
- INPUT for the PROMISE program is prepared, containing Np peaks defined with three coordinates, the position in 2D NOESY spectra (D1, D2) and the corresponding intensity *I*;
- The program calculates the distance matrix **D** with all peaks (Np, points contained in one set of peaks);
- For a chosen number of nearest neighbors (NN=1...Np), usually we chose NN = 1..6; a new matrix $\mathbf{D}^{NN}$ is made by retaining minimal distance elements from **D** in sequential steps, starting from one-element **D** matrix (1,1), ending with complete **D** matrix (Np,Np);
- Calculate row sums RS^{NN} of $\mathbf{D}^{NN}$ and calculate average row sum ARS^{NN};
- Derive the 2-component vectors of row sum RS^{NN} and intensity *I* for all rows in $\mathbf{D}^{NN}$. The 2-component vectors $V(RS^{NN},I)$ of dimension Np, which is the number of peaks in each set, are then evaluated for their average magnitude:

$$\mathrm{MV}^{\mathrm{NN}} = \frac{1}{\mathrm{Np}}\sum_{j=1}^{\mathrm{Np}} V_j^{\mathrm{NN}} = \frac{1}{\mathrm{Np}}\sum_{j=1}^{\mathrm{Np}} \sqrt{(\mathbf{RS}_j^{\mathrm{NN}})^2 + I_j^2} \quad (1)$$

The index *j* in eq 1 runs over all rows of the distance matrix $\mathbf{D}^{NN}$, from which the row sums $\mathbf{RS}_j^{NN}$ are calculated.

Usually we start the calculations for a limited number of peaks (*N*) from the set of Np peaks (*N*=10) and then enlarge

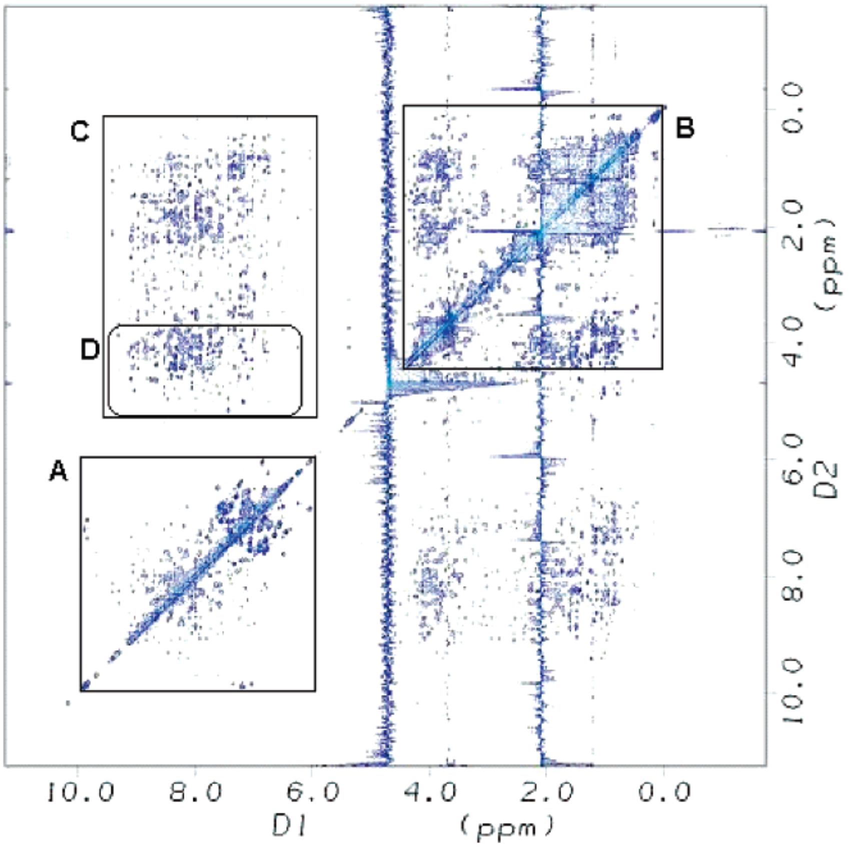

Figure 1. 2D NOESY spectrum of the protein isolate F1. Regions of typical NOE correlations are assigned: amide-aromatic (A), aliphatic (B), amide-aliphatic (C), and widths Hα-amide part (D). D1 and D2 are coordinates: D1, 6.0−10.2 ppm, D2, 1.1−5.4 ppm.

it to a few hundred. The distances involved in adding successive peaks, as *N* increases, are associated with shorter and shorter magnitudes. This is a consequence of the fact that the density of peaks in the spectral region continually increases, and thus new spots have neighbors at smaller separations.

3. RESULTS

3.1. Extraction of Spectral Data. 2D NMR NOESY spectra of four G-CSF isolates are shown in Figure 1.

In order to obtain information on protein structure from spectral data, four spectral regions that reflect the 3D structure of proteins were examined. In Figure 1 these regions are surrounded and labeled with letters A−D. Region A contains correlations assigned to the amide-aromatic part of the protein, B is typical for aliphatic, C for amide-aliphatic, and D for Hα-amide proton correlations. We have chosen the region C with amide-aliphatic proton correlations for further comparison, due to the high content of information about nonvicinal distances and the low number of artifacts usually present in this kind of spectra.

In Figure 2 one can observe a small difference in spectral region C (amide-aliphatic part) for the isolates G1 and F1.

3.2. Similarity Determination. Quantification of spectral similarity was based on the information content in the region C, which is characteristic for the amide-aliphatic part of the protein. The peaks above the noise level were extracted from spectra of individual isolates. The differences of spectral region C of different isolates were inspected in detail. At first glance the differences were very small; see Figure 2 for comparison of spectra of F1 and G1 isolates. However, when all peaks above the noise level were picked up, we obtained sets of about 1100 peaks for each isolate, which did not match completely. Two approaches were further investigated to obtain a reliable measure to quantitatively determine the degree of similarity. One is based on a direct statistical evaluation of matching peaks of compared spectra (only the peak positions were considered, not the intensities), while the other takes into account the distance matrix of a graph constructed on the basis of the most intense peaks with the sequential nearest neighbors method.[11]

3.2.1. Method 1. Method 1 is in fact a statistical evaluation of matching peaks. After the visual inspection of spectra there was a high similarity between isolates (Figure 2). Once the sets of peaks were compared, we found that there was a rather high number of nonmatching peaks. There are two reasons for such results: (i) nonmatching peaks of low intensity, which are visually overlooked and are presumably not very important, and (ii) a small difference in the position (D1, D2) of equivalent peaks in the spectra of two isolates. Thus we have to find the threshold, which would optimally eliminate low-intensity peaks (noise) and determine the tolerance of the peak position in both dimensions. We consequently calculated the amount of concurring peaks of the compared data sets of 2D NMR spectra for several thresholds (*P*) and for selected tolerances of the peak position in both spectral dimensions (T_{D1} and T_{D2}).

At constant tolerances ($T_{D1} = 0.01$ and $T_{D2} = 0.02$) we first varied the threshold above which the peaks were considered for comparison. Only a few best results out of 36 trials are shown in Table 1. The percentages of concurring peaks of isolates F1 and G1 are calculated as follows. If we

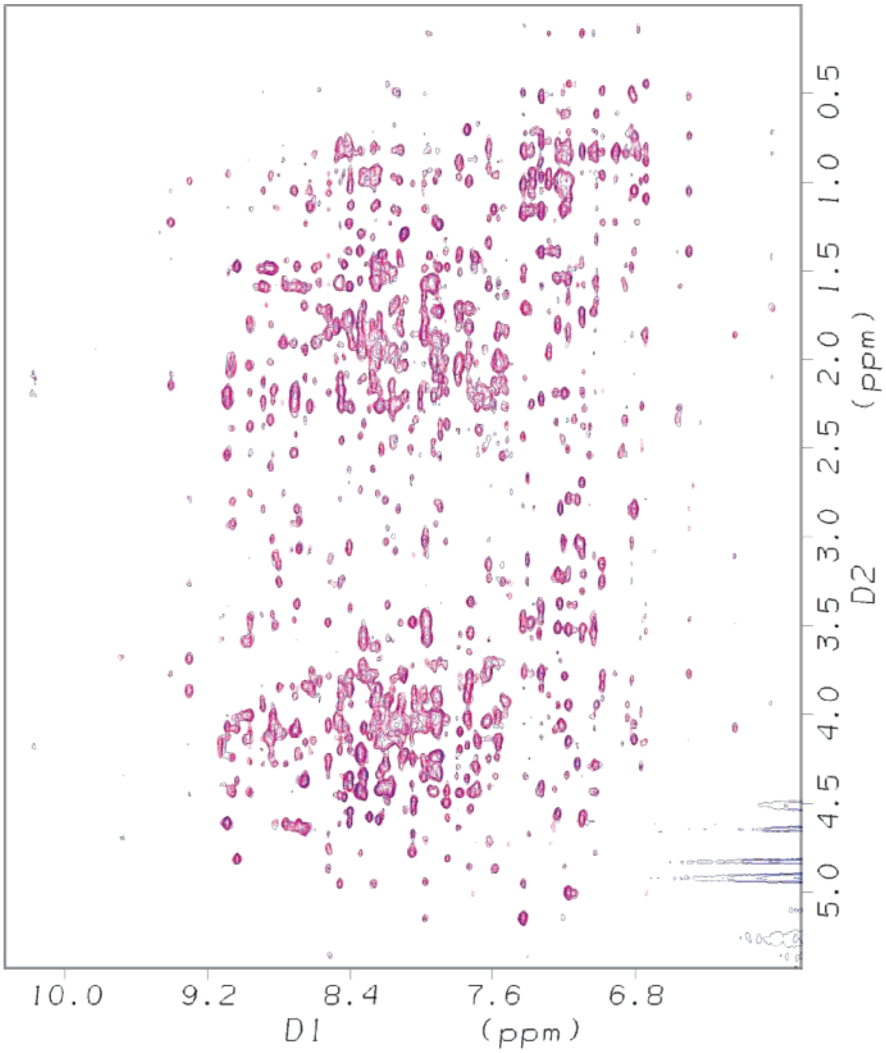

Figure 2. Overlapped NOESY spectra of the isolate F1 (blue) and G1 (red) in amide-aliphatic region C.

Table 1. Comparison of Isolates F1 and G1 for Several Threshold Values P^a

no.	P(F1)	P(G1)	no. of peaks F1	no. of peaks G1	all peaks (F1 or G1) $S^{G1} \cup S^{F1}$	common peaks $S^{G1} \cap S^{F1}$	matching peaks (%)
1	0	0	924	911	1103	732	66.36
2	0.005	0.005	921	908	1100	729	66.27
3	0.010	0	917	911	1099	729	66.33
...							
25	0.040	0.032	801	764	934	631	67.56
26	**0.040**	**0.033**	**801**	**759**	**930**	**630**	**67.74**
27	0.045	0.033	758	759	905	612	67.62
28	0.050	0.033	730	759	893	596	66.74
29	0.050	0.040	730	699	857	572	66.74
...							
33	0.041	0.033	795	759	927	627	67.64
34	0.050	0.050	730	620	828	522	63.04
35	0.060	0.050	651	620	771	500	64.85
36	0.060	0.060	651	548	738	461	62.47

[a] Sets of peaks S^{G1} and S^{F1} are obtained from the 2D NOESY spectral region C.

name the sets of peaks of the two isolates S^{G1} and S^{F1}, the intersection ($S^{G1} \cap S^{F1}$) defines the number of concurring peaks, while the union of these two sets ($S^{G1} \cup S^{F1}$) gives all possible peaks in any of the two sets. The percentage (column eight, Table 1) is then calculated as 100 × ($S^{G1} \cap S^{F1}$)/ ($S^{G1} \cup S^{F1}$). We can notice a low influence of the threshold on the percentage of matching peaks; obviously the initial peak-picking was performed above the noise level. The optimum was obtained at $P = 0.040$ and 0.033 for G1 and F1, respectively.

Tolerances T_{D1} and T_{D2} could not be optimized on the same way as the threshold intensity. In fact it has to be chosen in accordance with the expected error of peak position in both dimensions. Expected errors are 0.01 and 0.02 ppm in D1 and D2, respectively. Comparing the shifts of equivalent peaks in two spectra we found larger errors at more intense peaks. Consequently we decided to determine the tolerance relative to the peak widths W_{D1} and W_{D2}:

$$T_{D1} = W_{D1}/1.5 \text{ ppm} \tag{2}$$

$$T_{D2} = W_{D2}/1.5 \text{ ppm} \tag{3}$$

The resulting percentages of concurring peaks comparing all pairs of isolates (except for F1−F3) are collected in Table 2. We can see that the consideration of tolerances in both spectral dimensions improved the number of matching peaks for almost 10% in the comparison of F1 with G1. It is also obvious that F2 is the most different from all other isolates, while F1 and F3 are similar to G1, with over 70% of concurring peaks.

3.2.2. Method 2. Method 2 is based on graph invariants as described in section 2.3. Sets of peaks were used to

Table 2. Comparison of Pairs of Isolates F1, F2, F3, and G1, after Optimization for the Threshold (*P*1 and *P*2) and Chosen Tolerances (Eqs 2 and 3)

pair	*P*1	*P*2	all peaks[a] $S^1 \cup S^2$	common peaks[a] $S^1 \cap S^2$	common peaks (%) $(S^1 \cap S^2)/(S^1 \cup S^2)$
F1:G1	0.040	0.033	930	715	**76.88**
F1:F2	0.134	0.070	500	195	**39.00**
F1:F3	0.000	0.000	1262	847	**67.12**
F2:G1	0.090	0.070	551	231	**41.92**
F2:F3	0.269	0.304	149	71	**47.65**
F3:G1	0.000	0.014	1194	860	**72.03**

[a] S^1 and S^2 refer to a particular pair in column 1, i.e., S^{F1} and S^{G1} for the first line.

construct graphs. The $\mathbf{D}^{NN}$ matrices of graphs have been constructed on the basis of peaks in the 2D NMR NOESY spectra (C region) for all isolates of G-CSF in the study. In Figure 3 it is shown how the graph of the isolate G1 depends on the number of peaks (*N*) taking into account only one nearest neighbor (NN=1). The average row sums ARS^{NN} and average two-component vectors MV^{NN} (eq 1) have been calculated from corresponding distance matrices. Their dependence on the number of peaks considered is plotted in Figures 4 and 5. The calculations with a higher number of nearest neighbors showed only subtle differences. For comparison see the Supporting Information, Figures S1 and S2.

It is worthwhile to repeat that the graph invariants ARS^{NN} and MV^{NN} were calculated with the aim to determine the measure of similarity between several isolates. Having this in mind, we tested the sensitivity of the invariants, so that we determined them not only for the spectral region C but also for the region A, to see if they are able to differentiate between the two spectral regions of the same isolate. From the average row sums plots (Figure 4) we can observe that only at a very low number of peaks considered the differences were noticeable. On the contrary, the second invariant MV^{NN} (average magnitude of two-component vector) showed better sensitivity and differentiated well between the two spectral regions and also between the investigated isolates (Figure 5). One can observe in Figure 5 that the isolates F1, F2, and F3 represented by the MV^{NN} of spectral region C are grouped as three upper curves, G1 is below them, while the lowest curves belong to three isolates of spectral region A. It is obvious that the second invariant, the average magnitude of two-component vector MV^{NN}, is a better criterion in comparison of the isolates than the average row sums ARS^{NN}. We can also conclude that the number of nearest neighbors has only a subtle influence and for the sake of simplicity NN = 1 is the best choice for further applications.

3.2.3. Method 3. In order to make the two investigated methods described in 3.2.1 (peak matching) and 3.2.2 (sequential neighborhood) more comparable, we made a modification of the first method, so that the same restrictions were considered as in the second method: the number of peaks analyzed and inclusion of peak intensities in to criterion. First, we elevated the threshold to obtain only 100 peaks in each set. The results are shown in Table 3.

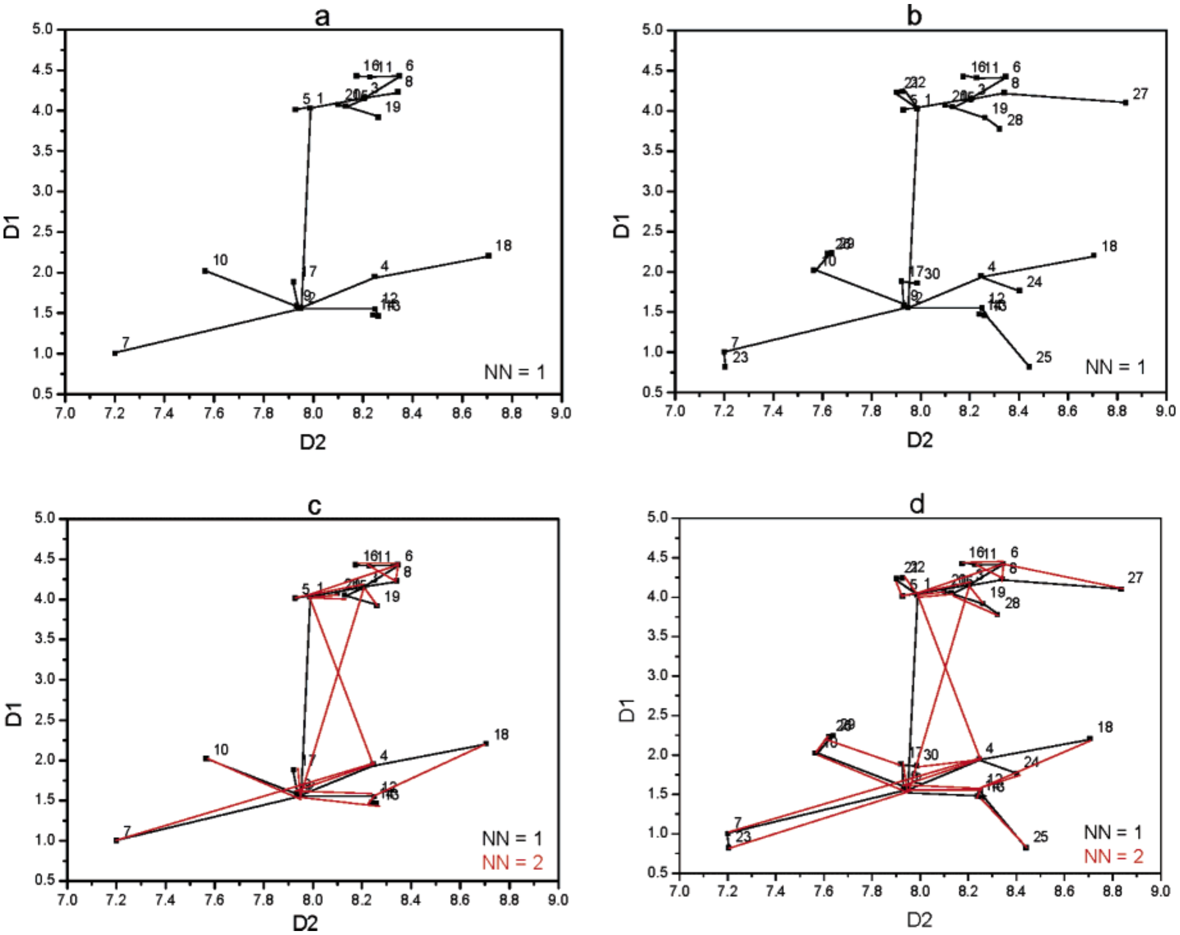

Figure 3. Graphs of connected peaks for G1 substrate, region C of 2D NOESY spectrum: (a) NN = 1, *N* = 20; (b) NN = 1, *N* = 30; (c) NN = 2, *N* = 20; and (d) NN = 2, *N* = 30.

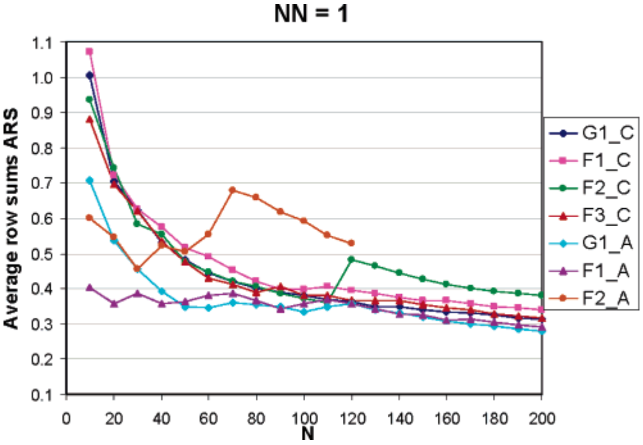

Figure 4. Average row sums ARS^{NN} versus number of peaks considered (N) for NN = 1.

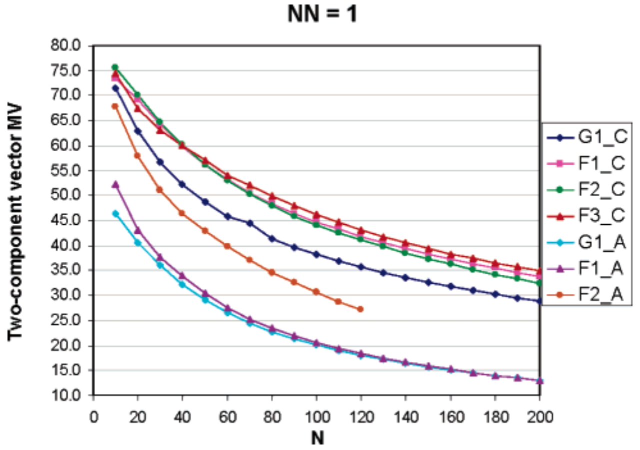

Figure 5. Dependence of the average magnitude of two-component vector MV^{NN} on the number of peaks considered (N) for NN = 1.

Table 3. Comparison of Pairs of Isolates F1, F2, F3, and G1, Considering Only 100 Peaks Obtained with Higher Threshold Intensity Values $P1$ and $P2$

			all	common	common	common	common
			peaks[a]	peaks[a]	peaks	peaks[b]	peaks[b]
pair	$P1$	$P2$	$S^1 \cup S^2$	$S^1 \cap S^2$	(%)	$\Sigma(1-\Delta I)$	(%)
F1:G1	0.290	0.245	121	93	**76.86**	82.07	**67.83**
F1:F2	0.290	0.287	164	64	**39.02**	54.67	**33.34**
F1:F3	0.290	0.286	128	84	**65.63**	75.30	**58.83**
F2:G1	0.287	0.245	165	61	**36.97**	52.41	**31.76**
F2:F3	0.287	0.286	154	66	**42.86**	58.28	**37.84**
F3:G1	0.286	0.245	124	87	**70.16**	76.14	**61.40**

[a] S^1 and S^2 refer to a particular pair in column 1, i.e., S^{F1} and S^{G1} for the first line. [b] Sum of common peaks weighted by the difference of peak intensities of compared pairs: $\Sigma(1-\Delta I)$.

From Table 3 we can see that the same pairs of isolates as in Table 2, obtained by the method 1, show the highest similarity (F1:G1 and F3:G1). Obviously the low-intensity peaks do not have a large influence on the comparison.

In the next step the peak intensities were included in the similarity determination of the isolates. The intensities of each set (isolate) were normalized to the maximal peak intensity. For each common peak (see Table 3, column 5) the absolute difference in the intensities was calculated, subtracted for 1 ($1-\Delta I$), and summed up for all common peaks in each pair of isolates (see Table 3, column 7). The largest contribution (i.e., 1) was obtained for common peaks of equal intensities; in case of a large intensity difference this contribution was even lower than 0.5. After division with the number of all peaks (Table 3, column 4) we obtain the percentage of matching peaks weighted by the difference in the intensity. Again, the same pairs of isolates were found the most similar, i.e., F1:G1 and F3:G1.

4. DISCUSSION

Two criteria for evaluation of similarity of 2D NMR NOESY spectra, direct peak-comparison (method 1), and graph theoretical invariants (method 2) described in sections 3.2.1 and 3.2.2, respectively, pinpointed different pairs of isolates as the most similar ones. Obviously different properties of the isolates are emphasized by the former criterion than by the latter one. The resulting percentages of concurring peaks (the first criterion) show the highest similarity between F1-G1 and F3-G1, with over 70% of concurring peaks. Also the pair F1−F3 shows a high degree of similarity (67%), while the isolate F2 is the most different from all other isolates. The comparison of the isolates by the graph invariant criterion demonstrated a higher similarity between the isolates F1, F2, and F3 than between G1 and any of the F isolates. The latter result seems quite reasonable taking into account that F1, F2, and F3 isolates were prepared by the same procedures (except for final formulation), while G1 as a commercially derived sample is produced by some other not exactly known procedure and formulated with a surfactant that cannot be completely removed by dialysis.

Searching for the reasons of the disagreement of the methods 1 and 2 we examined the essential influences:

- Method 1 takes into account **all** peaks from the selected spectral region (around 1000 for each isolate).
- In method 1 only the **position** of peaks are considered for comparison.
- Method 2 deals with a **limited number** of the most intense peaks from the selected spectral region (a few tens up to a few hundreds). The best selectivity was obtained at a low number of peaks considered (50−100).
- In method 2 the **intensity** is built in to the criterion not only implicitly by ordering peaks by the intensity but also directly in the two-component vectors MV^{NN}.

As described in the Results section, we introduced a modification of method 1 in order to make it more comparable to method 2 (see section 3.2.3, method 3). The threshold was elevated to the intensity value, which yielded 100 peaks per isolate. The similarities were found in the same order as before (method 1), which means that the low-intensity peaks did not contribute to the difference in the results obtained with the method 2. The same effect had an introduction of peak intensity difference considered as a weight in the summation of common peaks (see Table 3). Apparently the patterns of peaks considered in method 2 make the isolates F1 and F3 similar and distinguishable from G1. The expected difference of G1 is most likely due to the long-term stability additives. These might contribute to the high-intensity peaks and thus change the pattern of connections (see Figure 3) based on the intensity ordered peaks. A reason for differentiation of proteins could be in the quaternary structure, i.e., a potential agglomeration of the monomer protein; however, according to HPLC analytical data aggregation seems highly improbable in the case of F1−

F3 isolates. Additionally, all four isolates tested displayed the same biological activity, which is a strong indication for proper protein conformation. After all, additional experiments are needed to define the source of dissimilarity and give us a hint about the height of the proposed similarity criteria to be achieved for a new product, which would retain the desired biological property of the parent substrate.

5. CONCLUSIONS

Two methods were compared as a measure of similarity between the isolates of G-CSF. Both methods trace the information from the 2D NMR NOESY spectra in the region of the amide-aliphatic correlations of protons, which appear in the protein at a distance below 5 Å. The first method, which calculates the percentage of concurring peaks in the selected spectral region of two isolates, found the largest similarity between the isolates F1 and G1, while the second method between F1 and F2, with F3 close to them. As discussed above, the reason is not in a lower number of peaks considered in the second method, which was proved by a modification introduced by method 3. A restricted set of the most intense peaks gives a reasonable sensitivity of the obtained criterion, as can be appreciated from Figure 5. The differences in relative intensities change the sequence order of peaks considered in method 2 and thus influence the basic pattern of graph connections. Obviously this is the most important reason for obtaining different, however reasonable, similarities by method 2. We intend to continue the research with new experimental data, chosen with the aim to explain the source of differences in the proposed methods. One possibility is a point mutation of the protein at different (controllable) parts of the protein.

ACKNOWLEDGMENT

The financial support of this work by LEK Pharmaceuticals (BIO-7)) and by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia (Research Grant P1-017) is gratefully acknowledged. NMR spectra were recorded on spectrometers at the Institute for Biophysics Chemistry in Frankfurt, Germany (courtesy of Professor V. Doetsch).

Supporting Information Available: Behavior of the criterion calculated by method 2 for the higher number of nearest neighbors considered. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Šolmajer, T.; Zupan, J. Optimization algorithms and natural computing in drug discovery. *Drug Discovery today Technol.* **2001**, *1*, 247−252.

(2) Roberts, A. W. G-CSF: A key regulator of neutrophil production, but that's not all! *Growth Factors* **2005**, *23*, 33−41.

(3) Zink, T.; Ross, A.; Ambrosius, D.; Rudolph, R.; Holak, T. A. Secondary Structure of Human Granulocyte Colony-stimulating Factor Derived from NMR Spectroscopy. *FEBS Lett.* **1992**, *314*, 435−439.

(4) Hammerling, U.; Kroon, R.; Sjodin, I. In Vitro Bioassay With Enhanced Sensitivity for Human Granulocyte Colony-stimulating Factor. *J. Pharm. Biomed. Anal*. **1995**, *179*, 117−126.

(5) Souza, L. M.; Boone, T. C.; Gabrilove, J.; Lai, P. H.; Zsebo, K. M.; Murdock, D. C.; Chazin, V. R.; Bruszewski, J.; Lu, H.; Chen, K. K. Recombinant Human Granulocyte Colony-stimulating Factor: Effects on Normal and Leukemic Myeloid Cells. *Science* **1986**, *232*, 61−65.

(6) Jevševar, S.; Gaberc-Porekar, V.; Fonda, I.; Podobnik, B.; Grdadolnik, J.; Menart, V. Production of Nonclassical Inclusion Bodies from Which Correctly Folded Protein Can Be Extracted. *Biotechnol. Prog*. **2005**, *21*, 632−639.

(7) Wüthrich, K. *NMR of proteins and nucleic acids*; Wiley: New York, 1986; pp 117−129.

(8) Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of nuclear magnetic resonance in one and two dimensions*; Claredon Press: Oxford, 1987; pp 516−527.

(9) Pristovšek, P.; Franzoni, L. Stereospecific assignment of protein NMR resonances based on the tertiary structure and 2D/3D NOE data. *Inc. J. Comput. Chem*. **2006**, *27*, 791−797.

(10) West, B. D. *Introduction to Graph Theory*, 3rd ed.; Prentice Hall: 2001.

(11) Randić, M.; Novič, M.; Vračko, M. Novel Characterization of Proteomic Maps by Sequential Neighborhoods of Protein Spots. *J. Chem. Inf. Model.* **2005**, *45*, 1205−1213.

(12) FELIX software. http://www.accelrys.com (accessed March 5, 2007).

(13) König, D. *The Theory of Finite and Infinite Graphs*; Leipzig, 1936.

(14) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys*. **1947**, *15*, 419−420.

(15) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*. **1947**, *69*, 17−20.

(16) Randić, M. On graphical and numerical characterization of proteomics maps. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1330−1338.

(17) Randić, M.; Novič, M.; Vračko, M. On characterization of dose variations of 2-D proteomics maps by matrix invariants. *J. Proteome Res.* **2002**, *1*, 217−226.

(18) Randić, M.; Witzmann, F.; Vračko, M.; Basak, S. C. On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: application to peroxisome proliferators. *Med. Chem. Res.* **2001**, *10*, 456−479.

(19) Randić, M.; Zupan, J.; Novič, M. On 3-D graphical representation on proteomics maps and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339−1344.

(20) Randić, M.; Zupan, J.; Novič, M.; Gute, B. D.; Basak, S. C. Novel matrix invariants for characterization of changes of proteomics maps. *SAR QSAR Environ. Res.* **2002**, *13*, 689−703.

CI6005273