Screening of Matrix Metalloproteinases Available from the Protein Data Bank: Insights into Biological Functions, Domain Organization, and Zinc Binding Groups

Orazio Nicolotti, Teresa Fabiola Miscioscia, Francesco Leonetti, Giovanni Muncipinto, and Angelo Carotti*

Dipartimento Farmaco-Chimico, University of Bari, via Orabona 4, I-70125 Bari, Italy

Received April 4, 2007

A total of 142 matrix metalloproteinase (MMP) X-ray crystallographic structures were retrieved from the Protein Data Bank (PDB) and analyzed by an automated and efficient routine, developed in-house, with a series of bioinformatic tools. Highly informative heat maps and hierarchical clusterograms provided a reliable and comprehensive representation of the relationships existing among MMPs, enlarging and complementing the current knowledge in the field. Multiple sequence and structural alignments permitted better location and display of key MMP motifs and quantification of the residue consensus at each amino acid position in the most critical binding subsites of MMPs. The MMP active site consensus sequences, the C-α root-mean-square deviation (RMSd) analysis of diverse enzymatic subsites, and the examination of the chemical nature, binding topologies, and zinc binding groups (ZBGs) of ligands extracted from crystallographic complexes provided useful insights on the structural arrangements of the most potent MMP inhibitors.

INTRODUCTION

Matrix metalloproteinases (MMPs) are members of the large family of zinc- and calcium-dependent endopeptidases known as metzincins.[1] Subsequent to a long scientific debate among the biological, pharmacological, and medical communities, it has become clear that MMPs play a prominent role not only in the late stages of tumor progression, characterized by the degradation of the extracellular matrix, but also in the early phases of cancer development: MMPs have been shown to be involved in the regulation of several signaling pathways responsible for homeostatic regulation, innate immunity, release and cleavage of cell-surface receptors, and block of delivery of pro-apoptotic ligands. Beyond their well-known physiological roles in tissue turnover and remodeling, MMPs have been recently implicated in the release of angiostatin and endostatin (two natural angiogenesis blockers) and in the cleavage and inactivation of chemokines triggering organ-specific metastasis. As a result, the inhibition of MMPs might be useful in the early stages of cancer, i.e., before metastasis occurs, but be even deleterious in advanced stages as it might contrast host defense processes and normal tissue function. In this controversial scenario, MMPs have been classified in three large categories: (1) targets, when MMPs promote tumorigenesis requiring pharmacologic downregulation;[2] (2) anti-targets, when MMPs play essential roles in normal cells and, as a result, their inhibition can lead to undesirable side effects and/or even to disease progression; (3) countertargets, when MMPs do not play a relevant role in disease and their modulation may result in altered biological functions or side effects. Despite their ability to regulate quite diverse signaling pathways, MMPs are homologous enzymes with sequence identities and similarities ranging from 33% to 79% and from 50% to 88%, respectively. All MMPs contain a common sequence motif, HExGHxxGxxH, incorporating three histidines that coordinate the catalytic zinc ion, and typically share five-stranded β-sheets (one antiparallel and four parallel) and three α-helices in the zinc-based endopeptidase fold (Figure 1). Shaped as a cavity crossing the entire enzyme, the active site is characterized by a number of subsites directly involved in the interaction with physiological substrates and natural and synthetic inhibitors. The latter generally present a peptidomimetic moiety[3] linked to a functional group coordinating the catalytic zinc ion. On the basis of the catalytic domain[4] and some relevant structural/functional motifs (signal peptide, propeptide, convertase cleavage site, fibronectin type II modules, C-terminal, and so forth), MMPs have been divided into four main groups: (a) archetypal, (b) matrilysins, (c) gelatinases, and (d) convertase-activatable. Archetypal MMPs are organized around a catalytic domain constituted by a signal peptide, an amino-terminal propeptide, and a hemopexin carboxy-terminal domain relevant for substrate specificity and interactions with tissue MMP inhibitors. Archetypal MMPs include nine members (collagenases: MMP-1, MMP-8, MMP-13; stromelysins: MMP-3, MMP-10; metalloelastases: MMP-12; enamelysin: MMP-20; MMP-19; MMP-27). The second group is represented by matrilysins (MMP-7, MMP-26) in which the hemopexin C-terminal domain is absent and the catalytic domain is very small. Gelatinases (MMP-2 and MMP-9), belonging to the third group, contain three fibronectin type II modules structured as a thick collagen-binding domain that improve the degradation of collagen and gelatin. The presence of a basic insert in the propetide cleaved by furin-like proteases characterizes the fourth group composed of convertase-activatable MMPs. These proteins include three secreted MMPs, six membrane type (MT) MMPs, and two unusual type II transmembrane MMPs (MMP-23A, MMP-23B).

* Corresponding author e-mail: carotti@farmchim.uniba.it.

10.1021/ci700119r CCC: $37.00
Published on Web 10/24/2007

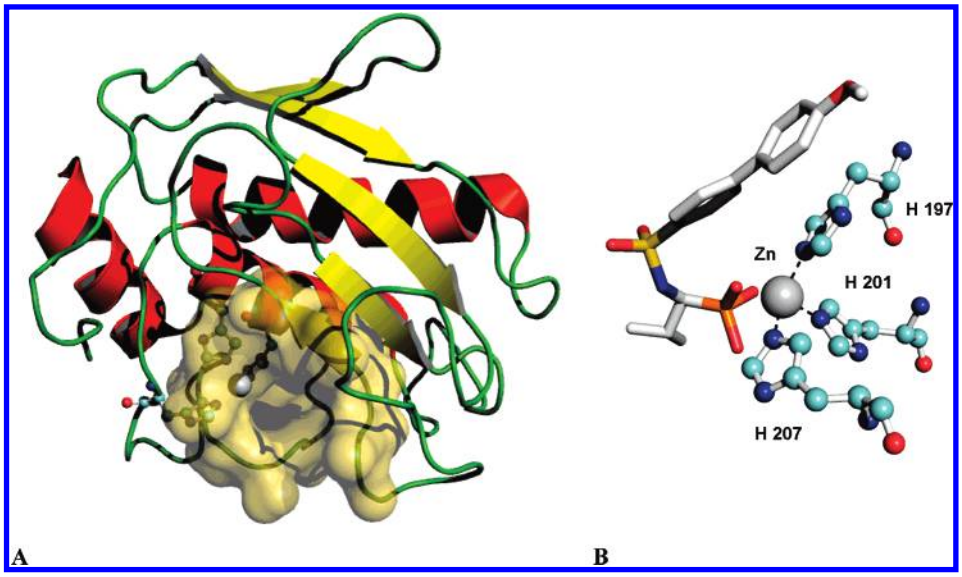

Figure 1. (A) Typical fold of matrix metalloproteinases structured in three α-helices (red) and four parallel and one antiparallel β-sheets (yellow). The binding site is represented by a white surface, while the zinc ion is shown as a light-gray sphere and the three catalytic histidines are rendered as ball-and-stick figures. (B) Coordination between zinc and the three catalytic histidines and the phosphonate group of the ligand extracted from 1ZS0 is highlighted.

More recent studies, consisting of multiple sequence alignments,[5] led to the elaboration of a complementary classification based on the analysis of phylogenetic trees. MMP-aligned sequences have been classified into four distinct groups: the glycosylphosphatidylinositol-anchored MMPs, transmembrane MMPs, gelatinases, and non-furin regulated MMPs. Other researchers[6] have performed a more extensive phylogenetic analysis on MMPs from all kingdoms. Such a phylogenetic tree analysis was made by including entire sequences in the multiple alignment and subsequently the C-terminal hemopexins and catalytic domains. Complementary analyses[7] have been performed on a low number (i.e., 10) of MMPs to evaluate MMP selectivity on the basis of GRID interaction molecular fields and consensus principal component analyses (CPCAs). Other studies,[8] addressing a higher number of MMPs (i.e., 24, including 15 structures from homology modeling), estimated the similarity within the MMP subsites by taking into account ligand interaction energies within the MMP subsites. Based again on GRID/CPCA, a further analysis[9] has been recently reported to evaluate MMP selectivity[10] on a larger number of proteins (i.e., 56 MMPs and 1 TACE).

Despite the wealth of functional information on MMPs,[11] a comparable amount of structural information has not yet been acquired. This observation prompted us to screen the entire Protein Data Bank (PDB) and take a census of all the available MMP three-dimensional structures. The ultimate purpose of the present work was the analysis and identification of new structural patterns[12] of MMPs potentially relevant for their classification and for the eventual design[13] of new classes of more potent and selective MMP inhibitors.

RESULTS AND DISCUSSION

For the sake of clarity, the entire methodological procedure described in the next paragraphs has been anticipated and summarized in the flowchart reported in Figure 2. As can be observed, the combined use of tools available from public servers and some bioinformatic routines developed in-house can be profitably applied to facilitate the analysis of any protein family well represented in the PDB.

NCBI Similarity Screening of MMPs Available in the PDB. Taken from recent literature, 12 MMP sequences from the PDB[14] (see marked PDB codes in Table 1) representing different MMP subfamilies were selected as queries for a similarity search on the National Center for Biotechnology Information (NCBI) server. The search pool was restricted to MMPs whose X-ray structures were taken from the PDB, since our aim was to derive structural insights potentially useful in the design of new classes of MMP inhibitors. PSI-BLAST[15] was used for running one-iteration similarity searching of human and nonhuman organisms for each selected query. After removing replicates and filtering out sequences other than metzincins (generally those with *E*-values lower than the standard threshold set in PSI-BLAST), a number of $n = 142$ sequences was retrieved (Table 1).

However, for the sake of clarity, it should be noted that some MMP sequences (i.e., MMP-15, MMP-19, MMP-20, MMP-21, MMP-23, MMP-24, MMP-26, and MMP-27) were not considered in our investigation since no structural information was available yet from the PDB. As will be seen later, this will not hamper the validity of our analysis. To better understand the sequence−function relationship of MMPs, pairwise sequence alignments[16] were automatically executed for all the $n(n-1)/2$ possible sequence combinations and score values, similarity, and identity were calculated using bioinformatic tools developed in our laboratory. Designed to find matches between small regions of amino acid sequences,[17] the Smith−Waterman algorithm was chosen to calculate local alignments[18] using point accepted mutation (PAM250) as a scoring matrix while the parameters gap-open and gap-extend were set to 10.0 and 0.5, respectively.

The alignment (Figure 3) relative to two collagenases of different classes was generated by *seqAA2simmat* (an in-house developed tool) that returned an excellent reproduc-

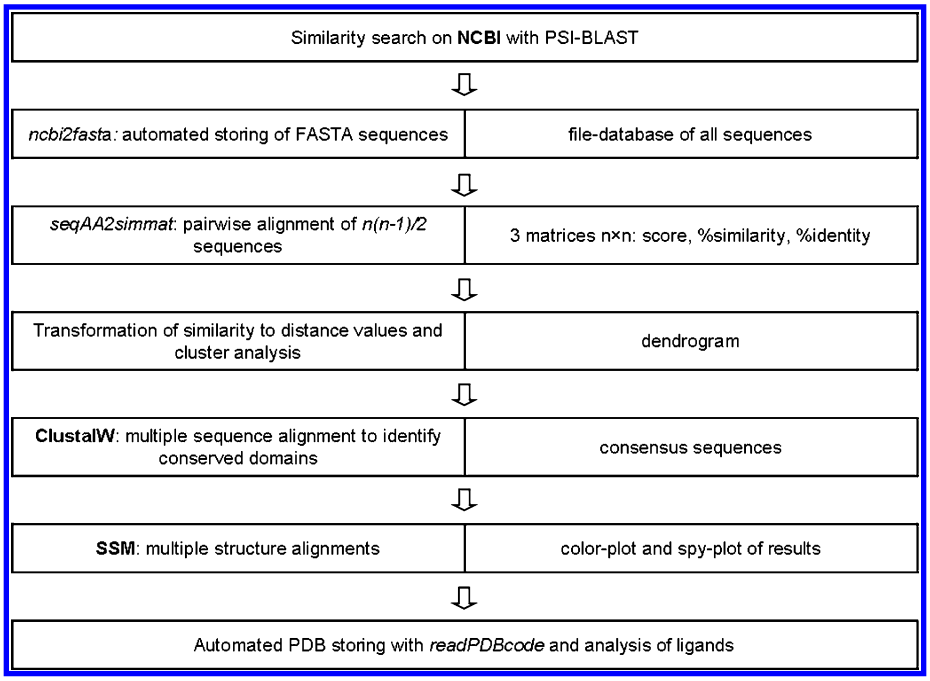

Figure 2. Flowchart showing the entire computational procedure for analysis of the MMPs.

Table 1. List of Matrix Metalloproteinase (MMP) Sequences Retrieved by PSI-BLAST for Which PDB Entries Are Also Available[a]

	subfamilies	classes	PDB codes
archetypal MMPs	collagenase	MMP-1	1CGE*, 966C, 1HFC, 4AYK, 3AYK, 2AYK, 1AYK, 2TCL, 1CGL, 1CGF, 2J0T
		MMP-8	1MNC, 1ZS0, 1ZP5, 1MMB, 1JAO, 1JAP, 1JAQ, 1JJ9, 1I76, 1I73, 1ZVX, 1BZS, 1KBC, 1JAN, 1A86, 1A85*, 1JH1
		MMP-13	1XUC, 1XUR, 1XUD, 1YOU, 830C, 456C, 1FM1, 1FLS, 1ZTQ, 2D1N, 1EUB, 1CXV*
	metalloelastase	MMP-12	1Z3J, 1YCM, 1Y93, 1RMZ, 1OS9, 1OS2, 1UTZ, 1UTT, 1JIZ, 1ROS, 1JK3*
	stromelysin	MMP-3	1B8Y, 1CIZ, 1CAQ, 3USN, 1OO9, 1G4K, 2USN, 1USN, 1SLM, 1UEA, 1HFS, 1QIC, 1C3I, 1CQR, 1BQO, 1BM6, 1BIW*, 1SLN, 1HY7, 1G05, 1G49, 1D5J, 1D8F, 1D7X, 1D8M, 2SRT, 2D1O, 1UMS, 1UMT, 1B3D, 1QIA, 1C8T
		MMP-10	1Q3A*
matrilysin	matrilysin	MMP-7	1MMP*, 1MMQ, 1MMR
gelatinases	gelatinase	MMP-9	1GKC, 1GKD
		MMP-2	1HOV, 1QIB
convertase-activatable MMPs	membrane-type MMP	MMP-14	1BUV, 1BQQ*
		MMP-16	1RM8*
	stromelysin-3	MMP-11	1HV5*
others	proenzyme	pro-MMP1	1FBL, 1SU3, 2CLT
		pro-MMP2	1CK7*, 1GXD, 1EAK
		pro-MMP9	1L6J
	additional single domain	fibronectin	1H8P, 1CXW, 2FN2, 1KS0, 1J7M, 1QO6, 1E88, 1E8B
		hemopexin	1ITV*, 1PEX, 1GEN, 1RTG, 1HXN
	MMP-related	ADAM	1R55, 1R54, 4AIG, 3AIG, 2AIG
		atrolysin C	1DTH
		snake venom metalloproteinase	1ND1, 1YP1, 1KUK, 1KUI, 1KUF, 1KUG, 1IAG
		TACE	2FV9, 2DDF, 2FV5, 2A8H, 1ZXC, 1BKC
		zinc protease	1KUH, 1C7K
		metzincins, methionine	1G08, 1G07
		serralysin	1SRP, 1SMP, 1AF0, 1SAT

[a] PSI-BLAST = Protein Specific Iterative−Basic Local Alignment Search Tool; PDB = Protein Data Bank. The asterisks indicate the sequences used as query for the PDB search.

ibility with results obtainable from the EMBOSS (European Molecular Biology Open Software Suite) server.[19]

Three 142 × 142 symmetric square matrices containing score, identity, and similarity values were obtained. In

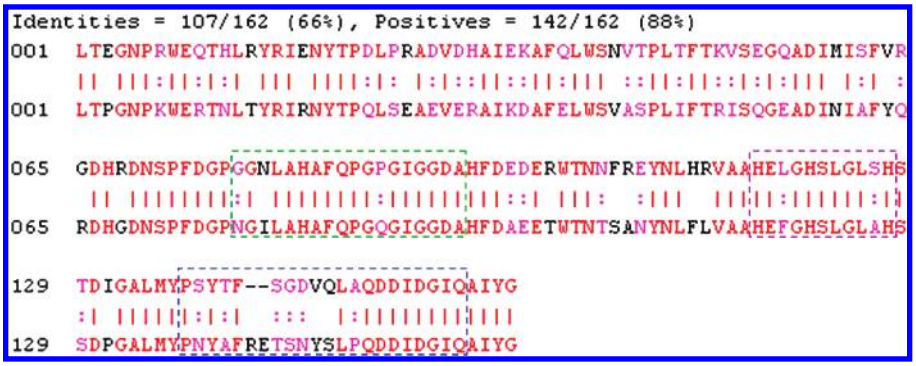

Figure 3. Pairwise sequence alignment relative to collagenases 1CGE (MMP-1) and 1ZS0 (MMP-8) obtained by *seqAA2simmat* (see text) using the Smith−Waterman algorithm with a point accepted mutation 250 scoring matrix. A score of 567 was calculated with identity and similarity being 107/162 (66%) and 142/162 (88%), respectively. Frames dotted in green (S_3', S_1, and S_3), violet (S_2 and S_2'), and blue (S_1') indicate the domains which are relevant for interaction with inhibitors (see text).

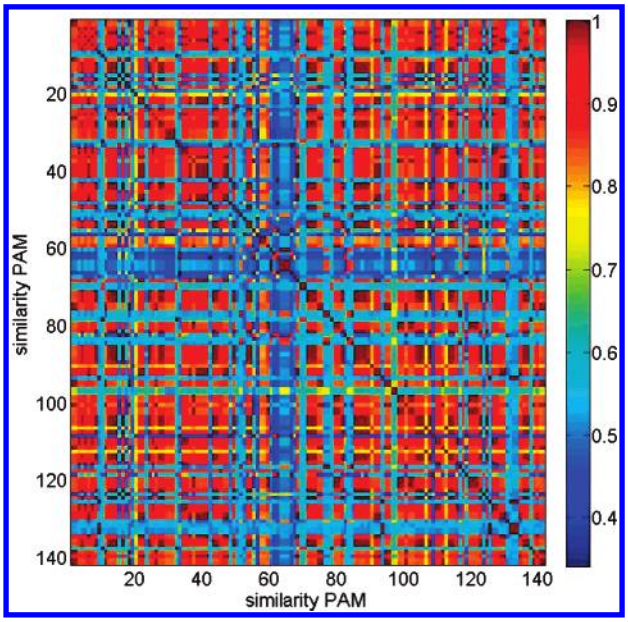

Figure 4. Heat maps relative to similarity values calculated for the $n(n-1)/2$ pairwise sequence combinations. Colors ranging from red to blue indicate shifts from high to low similarity values.

addition, the degrees of similarity calculated for each pair of sequences were displayed as heat maps in Figure 4, where a spectrum of colors ranging from blue to red indicates increment of pairwise similarity. The overwhelming occurrence of red areas in Figure 4 reveals a strong sequence similarity among MMPs and predicts a difficult and challenging design of highly selective ligands. The blue-shaded cross emerging from the plot indicates MMP sequences with a similarity less than 50%.

Cluster Analysis of MMP Similarity Data. To obtain a better quantitative idea of the sequence−function relation of MMPs, the similarity symmetric matrix was transformed into the corresponding distance symmetric matrix by applying the following simple equation:

$$d = \frac{1 - s}{s}$$

where s and d represent the similarity and distance values, respectively, for a given pair of amino acid sequences. A hierarchical approach was used to create cluster trees[20] in a multilevel scheme. Among the different statistical functions tested to determine sequence proximity, the McQuitty linkage (based on unweighted arithmetic averages and commonly implemented in the unweighted pair group method with arithmetic mean method for generating phylogenetic trees)[21] was selected for yielding the most interpretable dendrogram (Figure 5).

The dendrogram shows two main clusters joined at a similarity threshold of about 30%. Interestingly, the first cluster (located on the right-hand side of the graph) collected all the MMPs distributed according to the recently proposed domain organization classification.[4] Consequently, MMPs with similar motifs were grouped into the same cluster even if they belonged to different MMP subfamilies. From the dendrogram in Figure 5 one can observe that the domain classification[4] is clearly resolved on the right-hand side of the plot (yellow-shaded area), with the only exception of a group of gelatinases, MMP-2 (1HOV, 1QIB) and MMP-9 (1GKC, 1GKD) that to some extent are misplaced in the middle of archetypal MMPs.[22] This unexpected position might be due to the fact that gelatinases normally incorporate three fibronectin modules generally involved in specific binding with collagen. However, a closer analysis of MMP-2 and MMP-9 within the archetypal-MMP group reveals that such proteins lack the fibronectin modules which are substituted with the short peptide SLGKGV. Such a finding was significant given that such gelatinases (1HOV, 1QIB, 1GKC, and 1GKD) are very similar to archetypal MMPs and, at the same time, clearly different from matrilysins, which lack the hemopexin domain, and from convertase-activatable MMPs, which have an additional domain recognized by furin-like convertases. The left-hand side of the dendrogram contains MMPs with unique structural motifs (e.g., C-terminal hemopexin-like and fibronectin), pro-MMPs (MMPs integrating a propeptide fragment), and sequences denoted as MMP-related proteins. The latter can be considered as still belonging to the family of metzincins but, to some extent, cannot be properly regarded as MMPs. Provided with a lesser similarity to typical MMPs, MMP-related sequences belong to other classes of metzincin, such as the ADAM family (similar to snake-venom MMPs), which seem to be implicated in the regulation of migration and adhesion of tumor cells.[4] The cluster analysis disclosed other interesting details. MMPs of identical class were grouped into a unique subgroup with a similarity level of approximately 90% (Figure 6A). All these sequences belong to the MMP-3 class except for the green-colored sequence which is an

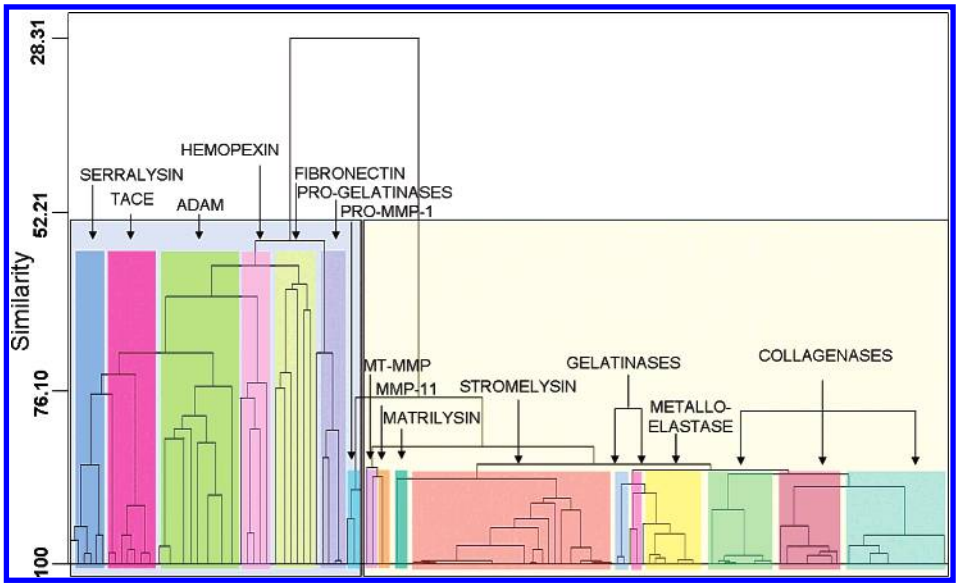

Figure 5. Cluster tree dendrogram plot of the 142 matrix metalloproteinase (MMP) proteins analyzed. MMP sequences and similarity values are reported on the horizontal and vertical axes, respectively. The links between sequences are represented as upside-down U-shaped lines with height indicating the corresponding percentage value of similarity.

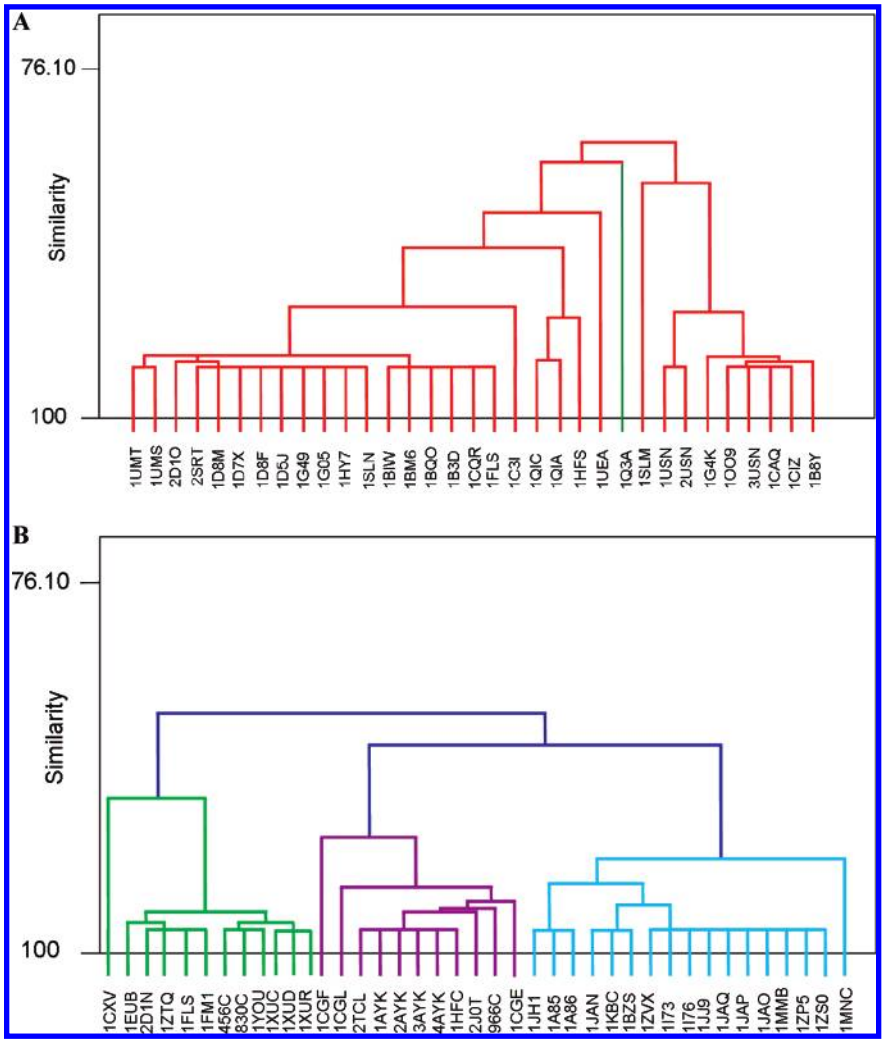

Figure 6. Closeup of Figure 5. Diagram A provides a detailed representation of matrix metalloproteinase MMP-3; diagram B illustrates the grouping of collagenases. The Protein Data Bank codes and similarity values (%) are reported on the horizontal and vertical axes, respectively.

MMP-10 and, to some extent, can be considered as a single-tone linking two MMP-3 subclusters. For a thorough understanding, it must be underlined that such a finding was consistent with recently reported phylogenetic trees showing these stromelysins grouped into a unique and well distinct cluster.[6] Different MMP classes were correctly joined to

Table 2. Analysis of Conserved Domains of Matrix Metalloproteinases (MMPs)[a] Relevant for the Ligand Interactions

	S_3'	S_1	S_3	catalytic domain ($S_2 + S_2'$)	S_1'
MMP-1	GGNL	AHA	FQPGPGIGGDA	HELGHSLGLSH	PSY-TF-SGDVQLAQ
MMP-2	DGLL	AHA	FAPGTGVGGDS	HEFGHAMGLEH	PIY-TYT-KNFRLSQ
MMP-3	GNVL	AHA	YAPGPGINGDA	HEIGHSLGLFH	PLYHSLTDLTRFRLSQ
MMP-7	GNTL	AHA	FAPGTGLGGDA	HELGHSLGMGH	PJY-GNGDPQNFKLSQ
MMP-8	NGIL	AHA	FQPGQGIGGDA	HEFGHSLGAH	PNY-AFRETSNYSLPQ
MMP-9	DGLL	AHA	FPPGPGIGGDA	HEFGHALGLAH	PMY-RFT-EGFPLHK
MMP-10	GHSL	AHA	YPPGPGLYGDI	HELGHSLGLFH	PLYNSFTELAQFRLSQ
MMP-11	GGIL	AHA	FFPKTHREGDV	HEFGHVLGLQH	PFY-TFR-YPLSLSP
MMP-12	GGIL	AHA	FGPGSGIGGDA	HEIGHSLGLGH	PTY-KYVDINTFRLSA
MMP-13	SGLL	AHA	FPPGPNYGGDA	HEIGHSLGLGH	PIY-KYTGKSHFMLPD
MMP-14	GGFL	AHA	YFPGPGIGGDT	HELGHALGLEH	PFY-QWMDTENFVLPD
MMP-16	GGFL	AHA	YFPGPGIGGDT	HELGHALGLEH	PFY-QYMETDNFVLPD
consensus	GGVL	AHA	FAPGPGIGGDA	HEIGHSLGLFH	PLYHSFTDLTRFRLSQ

[a] Enzyme subsites are indicated according to the conventional methods of Berger and Schechter (*Philos. Trans. R. Soc. London, B: Biol. Sci.* **1970**, *257*, 249−264).

higher level clusters which accurately reproduce the MMP subfamilies that were until now distinguished on the basis of their preference for a given substrate. For example, the three clusters relative to MMP-8, MMP-1, and MMP-13 merged to form a higher level and larger cluster populated by 40 sequences sharing a similarity value of approximately 85%. Such a cluster represents the well-known subfamily of collagenases (see Figure 6B). Similarly, the other remaining MMP subfamilies can be easily identified in the dendrogram.

Due to the continuously growing number of X-ray structures deposited into the PDB, our analysis must to be considered as an open eye on the PDB at the time of the beginning of our MMP investigation (December 2006). In fact, it should be observed that during the execution of this work nine new MMP X-ray human structures have been solved and made available from the PDB with the following codes: 2DDY (MMP-7), 2OY2 (MMP-8), 2OY4 (MMP-8), 2HU6 (MMP-12), 2OXU (MMP-12), 2OXW (MMP-12), 2OXZ (MMP-12), 2OW9 (MMP-13), and 2E2D (MMP-13). Although not initially considered, the inclusion of these fresh MMP sequences in our analysis did not change the clusterogram that interestingly collocated such new sequences where they were expected to be.

Multiple Sequence/Structure Alignment and Analysis of Ligands. To identify the conserved domains and, more importantly, the ones relevant for interacting with substrates and inhibitors, multiple alignment studies were conducted on 102 MMPs deriving from the main high-tree group (yellow-shaded area in Figure 5) and from the pro-MMP cluster (blue-shaded area) on the right-hand side of the dendrogram in Figure 5. The analysis was conducted by ClustalW[23] operating with default parameters. Jalview[24] was used to display results and to find conserved domains, which are those sequences having higher consensus.

By definition,[19] alignment consensus is the percentage of modal residue per column, and by default its calculation includes column gaps. Consensus analysis permitted the immediate identification of MMP binding site sequences that, as expected, were very well conserved across the entire MMP family. Sequences relative to diverse binding subsites of each MMP class, derived from multiple alignment, are reported in Table 2 along with the consensus calculated on a large number (102) of MMPs available from the PDB. To our satisfaction, perfect agreement became evident between the conserved domains in Table 2 and alignment, as reported in Figure 3 (see frames). The histograms indicating the percentage of consensus of each residue within a given subsite are shown in Figure 7.

To complement our sequence investigation, multiple structural alignment analysis was performed by submitting to the SSM (Secondary Structure Matching) server[25] the 142 X-ray MMP structures along with the option for the best conserved chain. For each given MMP pair, SSM returned C-α RMSd values along with *Q*-score parameters indicating the goodness of alignment and information relative to rotational and translational movements of initial C-α carbon positions. Stored in a square (142 × 142) matrix, RMSd values were displayed as a heat map in Figure 8A that revealed low and, more importantly, largely uniform RMSd values in the range between 0 and 1.5 Å apart for some fibronectin and hemopexin stretches. As expected, structures were much better conserved than sequences as clearly confirmed by comparing RMSd heat maps relative to X-ray structure and sequence similarity reported in Figures 8A and 4, respectively.

Additional analyses were undertaken to detect which protein sequences were less conserved with respect to their structures. To this end, the matrix difference, taken as an absolute value, between the RMSd values (normalized within the range 0−1) and the corresponding sequence distances was calculated and further analyzed. Dropping to zero difference values lower than a given threshold (for convenience set to 0.65 in the present study), it resulted (Figure 8B) that pro-gelatinases only showed relevant differences between structure and sequence. In most cases pro-gelatinases were quite dissimilar from canonical MMPs but close to fibronectins, MMP-related, and pro-MMP1. Such results have already been displayed in the clusterogram (Figure 5), indicating that pro-gelatinases were first linked to fibronectins (as they are the only ones including their sequences), then to MMP-related and still to pro-MMP1 (sharing with them a common propeptide), and finally to all the remaining MMPs.

Finally, additional structural alignments were performed on the pool of the 102 MMPs by maximizing first the superposition only of C-α belonging to the subsites relevant for binding and then considering all the C-α of the MMP backbone. By computing the RMSd differences, taken as absolute values, between subsites and backbone C-α,

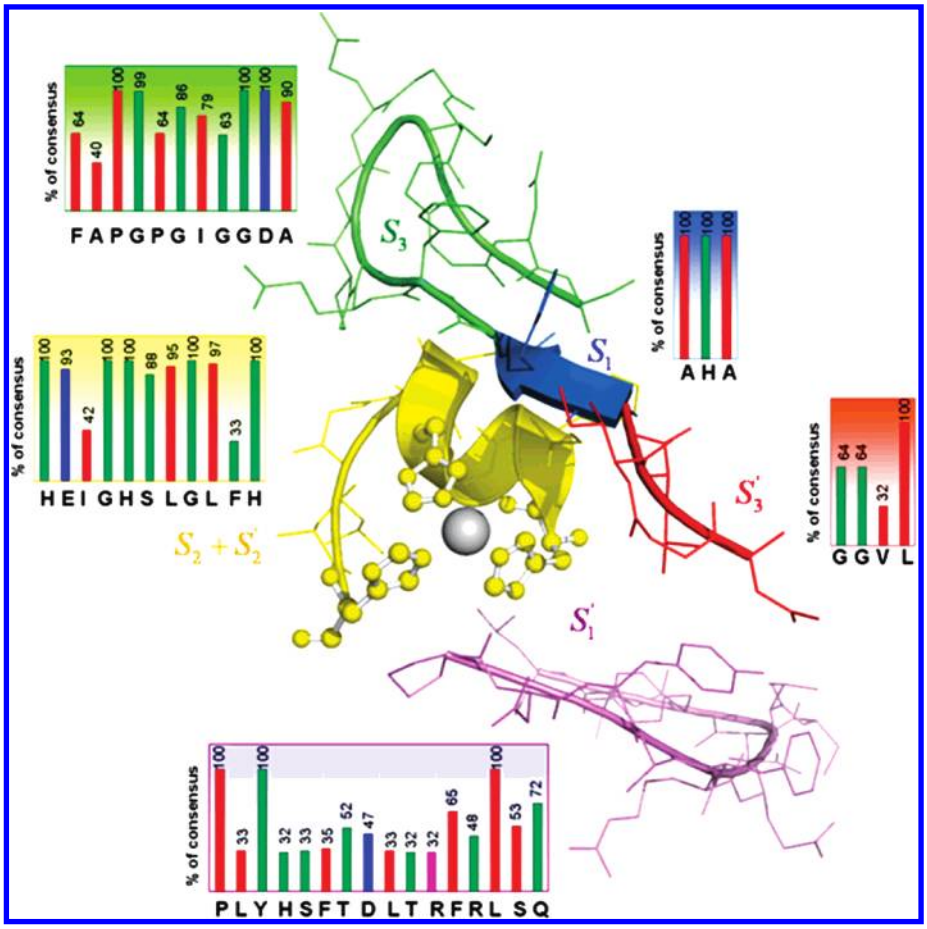

Figure 7. Display of domain motifs together with their corresponding histograms showing the percentage consensus at the indicated residue positions.

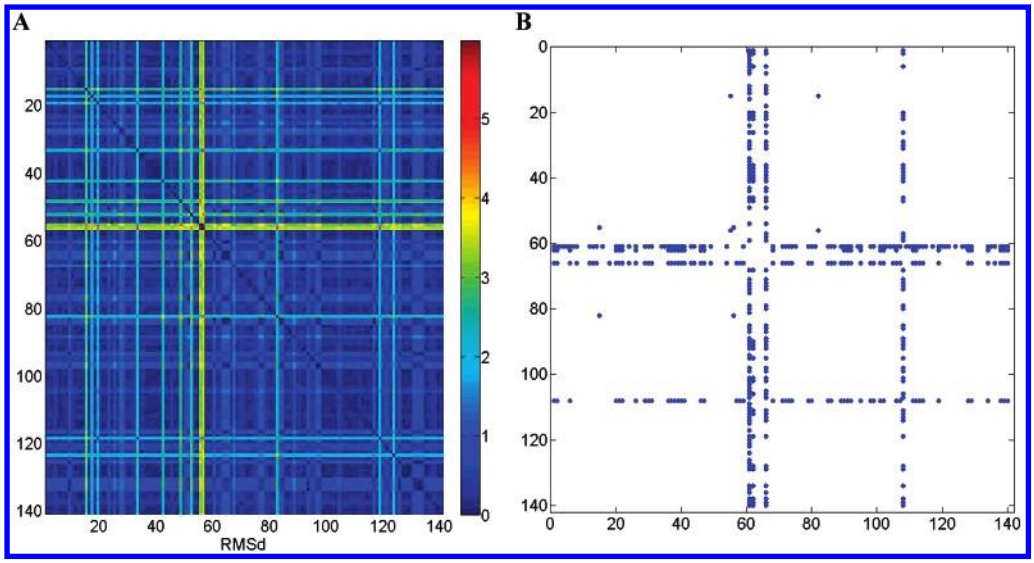

Figure 8. (A) Heat maps relative to RMSd values calculated for the $n(n-1)/2$ pairwise structural combinations. (B) Spy plot relative to the difference, taken as an absolute value, between the normalized (0−1) RMSd structure-based and distance sequence-based matrices, respectively. Absolute differences lower than 0.65 were dropped to zero.

it resulted (Figure 9A) that MMP-11 (1HV5, stromelysin-3) was the only X-ray structure significantly different from all the others with an average RMSd equal to 2.7 Å. As an example, a poor structural overlap (RMSd = 3.003 Å) at the binding subsites was observed between 1HV5 and 1HY7 (MMP-3, stomelysin-1). Interestingly, a simple visual inspection revealed that such a pronounced RMSd value was basically due to the remarkable structural difference occurring at subsite S3 (Figure 9B).

On the basis of these sequence/structure alignments, all the corresponding X-ray protein structures were superimposed onto the Cartesian coordinates of the high-resolution X-ray structure of 1ZS0, arbitrarily chosen as MMP template. A number of 92 chemical ligands were extracted accordingly

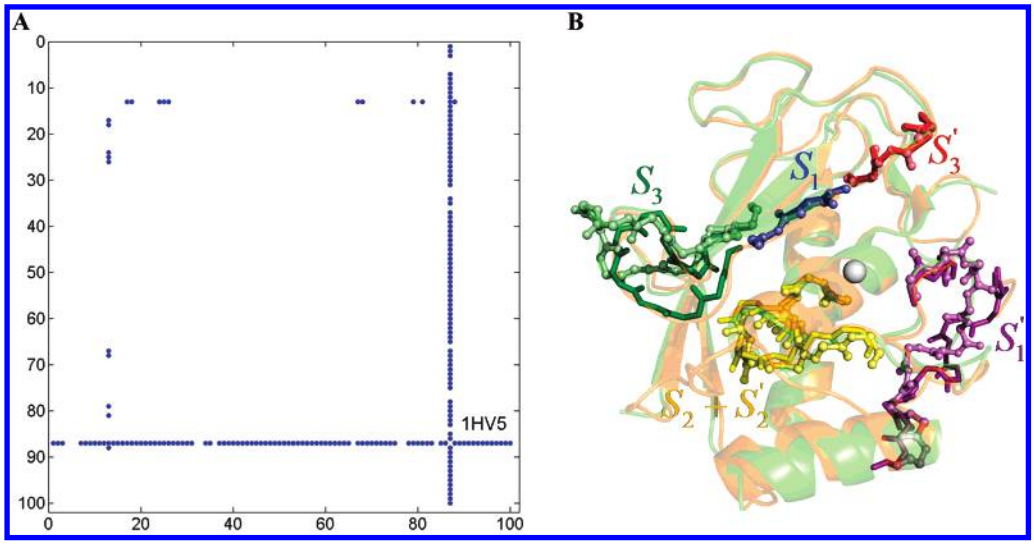

Figure 9. (A) Spy plot relative to the difference, taken as an absolute value, between RMSd values calculated on C-α of the backbone and the binding subsites, respectively. RMSd values lower than 1 were dropped to zero. (B) 1HV5 and 1HY7 structural superposition rendered as ball-and-stick lighter and capped-stick darker tints, respectively. Stick and cartoon models are referred to binding subsites and backbone, respectively. Subsite S_3 shows remarkable structural divergence (see text).

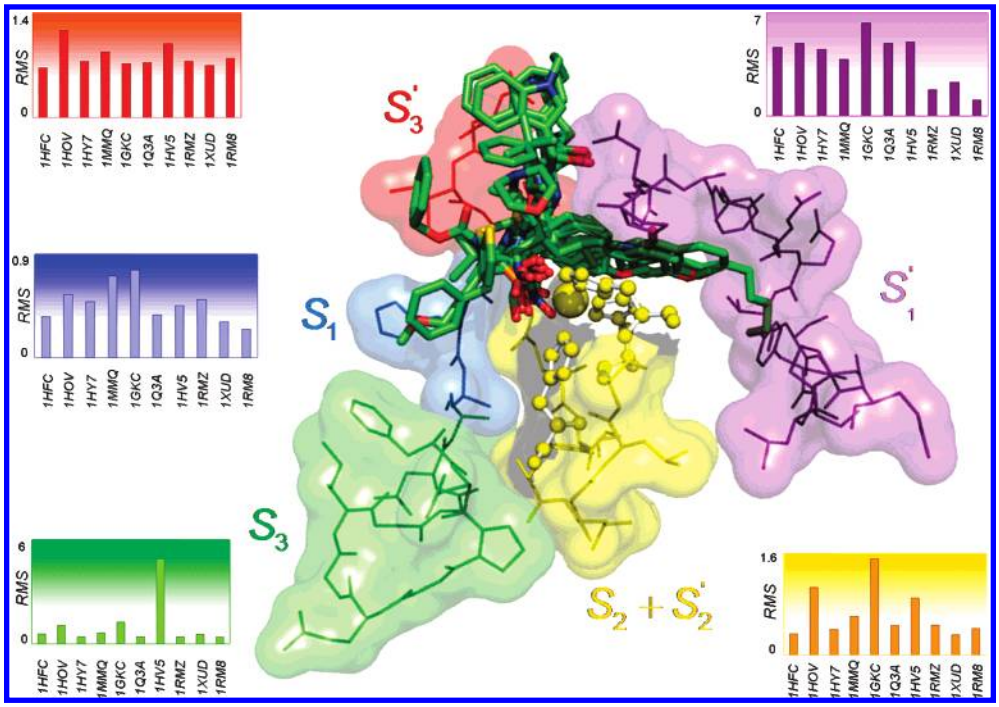

Figure 10. Complex of 1ZS0 with all X-ray ligands extracted from the central matrix metalloproteinase model. Computed on the C-α of binding subsites, histograms represent the RMSd values (Å) with respect to 1ZS0 chosen as a template.

by retrieving the coordinates of the non-hydrogen atoms from the original PDB files. Zinc binding groups (ZBGs) were unevenly distributed, being constituted by 53 hydroxamates, 20 carboxylates, six thiols, five phosphonates, three barbiturates, and five miscellaneous chemical classes. As is well-known,[26] such ZBGs engage monodentate or bidentate coordination with zinc exhibiting tetrahedral or trigonal bipyramidal geometry, respectively. Visual inspection of the ZBGs revealed the extent of similarity of spatial arrangement adopted by the diverse ligands. For each MMP class, a protein was chosen as a model on the basis of high X-ray crystallographic resolution and the presence of a tightly bound inhibitor. The following MMP-inhibitor complexes were selected: 1HFC (MMP-1), 1HOV (MMP-2), 1HY7 (MMP-3), 1MMP (MMP-7), 1ZS0 (MMP-8), 1GKC (MMP-9), 1Q3A (MMP-10), 1HV5 (MMP-11), 1RMZ (MMP-12), 1XUD (MMP-13), and 1RM8 (MMP-16). MMP-14 was not considered given the lack of the inhibitor in the three-dimensional crystallographic structure. After centering the Cartesian coordinates of each protein on those of 1ZS0, the RMSd values, expressed in angstroms, were calculated taking into account the C-α overlay of each subsite sequence.[27] Higher RMSd values were observed for the S_1' subsites that constituted a critical protein pocket (violet-shaded area), generally binding the more hindered moieties of the inhibitor.[28] Such a finding suggested the important role of the S_1' subsite structure in the regulation of MMP selectivity.[9] Though not yet supported by further proof, a high RMSd

value (4.88 Å) was observed for the S_3 subsite in 1HV5 as already observed when conducting structural analysis. Interestingly, the difficult design of selective MMP inhibitors might in part arise from the high structural similarity (low RMSd values) measured for all the remaining MMP subsites. To simplify interpretation, the structural superimposition of the cocrystallized ligands of MMP selected as a model is reported in Figure 10. As can be observed, the hydrophobic and generally bulkier molecular moieties were embedded in the S_1' violet pocket. The peptidomimetic moieties engaged additional interactions into the regions delimited by the S_3, S_1, and S_3' subsites (green-, blue-, and red-shaded areas, respectively), while ZBGs faced the metal ion.

Computational Methods. A number of Matlab-based[29] bioinformatic tools were developed in our laboratory with the aim of speeding up and automating the entire process of protein search and analysis. Drawing from an extensive list of NCBI entries, protein sequences were rapidly acquired and stored in FASTA format by browsing the NCBI server with our routine *ncbi2fasta*. Full enumerated combinations[30] (i.e., $n(n-1)/2$) of pairwise comparisons were therefore executed on retrieved sequences by running our *seqAA2simmat* routine that returned three symmetric square matrices containing score, similarity, and identity values, respectively. The *seqAA2simmat* function offers options for selecting both Needleman−Wunsch[31] and Smith−Waterman[32] algorithms[33,34] to perform global and local matches, respectively. PAM, BLOSUM, and GONNET are available in *seqAA2simmat* as scoring matrices,[35] and additional controls such as the gap-open, gap-extend, and scale value are also user-definable. In addition, such matrices can be transformed into color plots, here referred to as heat maps, to display the global amount of given information within the data. In the present work, the Smith−Waterman algorithm was used to run sequence alignments by selecting the PAM250 scoring matrix and setting gap-open, gap-extend, and scale value at 10.0, 0.5 and 3.0, respectively. Score and identity values were consistent with EMBOSS results. Indeed, although negligible, slightly improved similarity values were calculated by *seqAA2simmat*, which included not only scoring matrix values greater than zero but also those equal to zero according to the well-known definition of sequence similarity. After converting similarity to the distance matrix as explained above, the statistical software Minitab[36] was used to run cluster analyses and to display dendrogram plots. ClustalW[23] was used for multiple sequence alignment, and Jalview[24] was adopted for displaying results. SSM server was utilized to carry out multiple structural alignments. The list of MMP sequences available from the PDB was subsequently automatically downloaded by running the in-house tool *readPDBcode*. Protein and ligand manipulations were executed within Sybyl.[37] Pictures were elaborated with Open-Source PyMOL.[38]

CONCLUSIONS

The main objective of the present investigation was to take a census, and thus to obtain a comprehensive overview, of all the MMP sequences for which X-ray structures were available. After retrieving a total of 142 MMP structures from the PDB, through PSI-BLAST, heat maps and hierarchical clusterograms were generated by running a series of bioinformatic tools developed in-house. Based on a far larger number of MMP sequences than ever used before and on calculations accounting for primary sequences, our analysis afforded a more reliable and comprehensive representation of the relationships existing among MMPs. Interesting outcomes also emerged from multiple sequence alignment which permitted better location and display of the MMP motifs and, more importantly, quantification of the residue consensus at each amino acid position of the most critical binding subsites. In addition, the comparison between sequence distance matrix and RMSd values, obtained through multiple structural alignment on the 142 targets, permitted us to quickly identify the more diverse proteins. Interestingly, our investigation resulted that differences were more pronounced, as can be seen for the family of pro-gelatinases, in the protein sequences rather than in structures that were, as expected, better conserved. Focusing the multiple structural analysis to the 102 canonical MMPs, it was easier to investigate spatial divergences by mainly considering those depending on the active site. The investigation on the MMP ligands extracted from protein−ligand complexes and their structural superimposition provided insights into structural features and binding modes of potent MMP inhibitors. In particular, the results from both consensus analysis and C-α RMSd computation indicated the S_1' subsite as the one playing a pivotal role in determining selectivity as unequivocally evidenced from the up-and-down profile (Figure 7) of the percentage of consensus and the wide variation of the RMSd values (up to approximately 7 Å, Figure 10). Taken together, the main findings of our investigation may help the design[39] of new MMP inhibitors with both improved potency and, hopefully, selectivity.

ACKNOWLEDGMENT

The authors wish to thank the students of MDM-Lab (Molecular & Data Modelling Laboratory) at the University of Bari for an appreciated collaboration. Credit is due to Dr. Valeria De Marco of the Division of Molecular Carcinogenesis at The Netherlands Cancer Institute for her helpful comments. Finally, the authors gratefully acknowledge the European Commission ("CancerGrid" STREP project, FP VI, Contract No. LSHC-CT-2006-03755), MIUR (Rome, Italy; PNR Project No. RBNE01F5WT), and Regione Puglia (Strategic Project "NeuroBiotech") for financial support.

Supporting Information Available: Chemical structures of 92 molecules extracted from crystallographic complexes of metzincins. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Burzlaff, N. Model Complexes for Zinc-Containing Enzymes. *Concepts Models Bioinorg. Chem.* **2006**, *17*, 397−429.
(2) Overall, C. M.; Kleifeld, O. Tumor Microenvironment−Opinion: Validating Matrix Metalloproteinases as Drug Targets and Anti-Targets for Cancer Therapy. *Nat. Rev. Cancer* **2006**, *6*, 227−239.
(3) Baran, I.; Varekova, R. S.; Parthasarathi, L.; Suchomel, S.; Casey, F.; Shields, D. C. Identification of Potential Small Molecule Peptidomimetics Similar to Motifs in Proteins. *J. Chem. Inf. Model.* **2007**, *47*, 464−474.
(4) Overall, C. M.; Lopez-Otin, C. Strategies for MMP Inhibition in Cancer: Innovations for the Post-Trial Era. *Nat. Rev. Cancer* **2002**, *2*, 657−672.
(5) Andreini, C.; Banci, L.; Bertini, I.; Luchinat, C.; Rosato, A. Bioinformatic Comparison of Structures and Homology−Models of Matrix Metalloproteinases. *J. Proteome Res.* **2004**, *3*, 21−31.

(6) Massova, I.; Kotra, L. P.; Fridman, R.; Mobashery, S. Matrix Metalloproteinsases: Structure, Evolution and Diversification. *FASEB J.* **1998**, *12*, 1075−1095.
(7) Terp, G. E.; Cruciani, G.; Christensen, I. T.; Jorgensen, F. S. Structural Differences of Matrix Metalloproteinases with Potential Implication for Selectivity Examined by the GRID/CPCA Approach. *J. Med. Chem.* **2002**, *45*, 675−2684.
(8) Lukacova, V.; Zhang, Y.; Mackov, M.; Baricic, P.; Raha, S.; Calva, J. A.; Balaz, S. Similarity of Binding Sites of Human Matrix Metalloproteinases. *J. Biol. Chem.* **2004**, *279*, 14194−14200.
(9) Pirard, B.; Matter, H. Matrix Metalloproteinase Target Family Landscape: a Chemometrical Approach to Ligand Selectivity Based on Protein Binding Site Analysis. *J. Med. Chem.* **2006**, *49*, 51−69.
(10) Matter, H.; Schwab, W. Affinity and Selectivity of Matrix Metalloproteinase Inhibitors: a Chemometrical Study from the Perspective of Ligands and Proteins. *J. Med. Chem.* **1999**, *42*, 4506−4523.
(11) Amin, E. A.; Welsh, W. J. A Preliminary in Silico Lead Series of 2-Phthalimidinoglutaric Acid Analogues Designed as MMP-3 Inhibitors. *J. Chem. Inf. Model.* **2006**, *46*, 2104−2109.
(12) Verma, R.; Hansch, C. Matrix Metalloproteinases (MMPs): Chemical-Biological Functions and (Q)SARs. *Bioorg. Med. Chem.* **2007**, *15*, 2223−2268.
(13) Nicolotti, O.; Gillet, V. J.; Fleming, P.; Green, D. A Novel Approach to Deriving Accurate and Chemically Intuitive QSAR Models. *J. Med. Chem.* **2002**, *45*, 5069−5080.
(14) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(15) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(16) Friedberg, I.; Harder, T.; Kolodny, R.; Sitbon, E.; Li, Z.; Godzik, A. Using an Alignment of Fragment Strings for Comparing Protein Structures. *Bioinformatics* **2006**, *23*, 219−224.
(17) Pearson, W. R. Searching Protein Sequence Libraries: Comparison of the Sensitivity and Selectivity of the Smith-Waterman and FASTA Algorithms. *Genomics* **1991**, *11*, 635−650.
(18) Altschul, S. F.; Gish, W. Local Alignment Statistics. *Methods Enzymol.* **1996**, *266*, 460−480.
(19) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.
(20) Nikolski, M.; Sherman, D. J. Family Relationships: Should Consensus Reign? Consensus-Clustering for Protein Families. *Bioinformatics* **2007**, *23*, 71−76.
(21) Whelan, S.; Lio, P.; Goldman, N.; Molecular Phylogenetics: State-of-the-Art Methods for Looking into the Past. *Trends Genet.* **2001**, *17*, 262−272.
(22) Steffensen, B.; Bigg, H. F.; Overall, C. M. The Involvement of the Fibronectin Type II-Like Modules of Human Gelatinase A in Cell Surface Localization and Activation. *J. Biol. Chem.* **1998**, *273*, 20622−20628.
(23) Thompson, J. D.; Higgins, D. G.; Gibbson, T. J. ClustalW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.
(24) Clamp, M.; Cuff, J.; Searle, S. M.; Barton, G. J. The Jalview Java Alignment Editor. *Bionformatics* **2004**, *20*, 426−427.
(25) Krissinel, E.; Henrick, K.; Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr.* **2004**, *60*, 2256−2268.
(26) Overall, C. M.; Kleifeld, O. Towards Third Generation Matrix Metalloproteinase Inhibitors for Cancer Therapy. *Br. J. Cancer* **2006**, *94*, 941−946.
(27) Najmanovich, J. R.; Allali-Hassani, A.; Morris, J. R.; Dombrovsky, L.; Pan, P. W.; Vedadi, M.; Plotnikov, A. N.; Edwards, A.; Arrowsmith, C.; Thornthon, J. M. Analysis of Binding Site Similarity, Small-Molecule Similarity and Experimental Binding Profiles in the Human Cytosolic Sulfotransferase Family. *Bioinformatics* **2007**, *23*, 104−109.
(28) Amin, E. A.; Welsh, W. J. Highly Predictive CoMFA and CoMSIA Models for Two Series of Stromelysin-1 (MMP-3) Inhibitors Elucidate S1′ and S1−S2′ Binding Modes. *J. Chem. Inf. Model.* **2006**, *46*, 1775−1783.
(29) *MATLAB The Language Of Technical Computing*, version 7.3; The Mathworks: Natick, MA, 2006.
(30) Nicolotti, O.; Carotti, A. QSAR and QSPR Studies of a Highly Structured Physico-Chemical Domain. *J. Chem. Inf. Model.* **2006**, *46*, 264−276.
(31) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(32) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.
(33) Marko, A. C.; Stafford, K.; Wymore, T. Stochastic pairwise alignment and scoring methods for comparative protein structure modelling. *J. Chem. Inf. Model.* **2007**, *47*, 1263−1270.
(34) Kratochwil, N. A.; Malherbe, P.; Lindermann, L.; Ebeling, M.; Hoener, M. C.; Mühlemann, A.; Porter, R. H. P.; Stahl, M.; Gerber, P. R. An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application *J. Chem. Inf. Model.* **2005**, *45*, 1324−1336.
(35) Altschul, S. F. Amino Acid Substitution Matrices from an Information Theoretic Perspective. *J. Mol. Biol.* **1991**, *219*, 555−565.
(36) Minitab Inc., 13.20; Making Data Analysis Easier, 2004.
(37) *SYBYL 7.1*; Tripos Inc.: 1699 South Hanley Road, St. Louis, MO 63144, 2007.
(38) *Pymol 0.99*; Delano Scientific LLC: South San Francisco, CA, 2006.
(39) Catto, M.; Nicolotti, O.; Leonetti, F.; Carotti, A.; Favia, A.; Soto-Otero, R.; Mendez-Alvarez, E.; Carotti, A. Modulation of Monoamine Oxidase-A and -B Inhibitory Activities of Coumarins by Systematic Substituents Variation at Position 7. *J. Med. Chem.* **2006**, *16*, 4912−4925.

CI700119R