Estendendo Consultas Contínuas por Abrangência aos Dados Métricos

Enzo Seraphim[1], Thatyana F. P. Seraphim[1], Lucio F. D. Santos[2], Edmilson Marmo Moreira[1], Luiz Olmes Carvalho[1]

[1]Universidade Federal de Itajubá (UNIFEI) – Itajubá (MG), Brasil

[2]Instituto Federal do Norte de Minas Gerais (IFNMG) – Montes Claros (MG), Brasil

{seraphim, thatyana, edmarmo, olmes}@unifei.edu.br, lucio.santos@ifnmg.edu.br

Abstract. *Several studies highlight the importance and complexity of determining the nearest points along a trajectory through a technique known as Continuous Query. Existing solutions for answering such queries rely on Euclidean coordinates in spatial data to verify whether the object of interest lies within the coverage of the route. However, this approach does not apply to purely metric data, as such data cannot be represented in an n-dimensional space. On the other hand, a distance function can be used to infer line segments over the metric space, similar to the association of a geometry. In this context, the present work associates Euclidean properties with the metric space to define the relationship between points and lines, extending continuous range queries to this data domain. The experiments demonstrate that this association enables efficient continuous range query on purely metric data, addressing the gap left by existing techniques in the literature.*

Resumo. *Diversos estudos mostram a importância e a complexidade para determinar os pontos mais próximos de uma trajetória, através de uma técnica conhecida como Consulta Contínua. As soluções existentes para responder essas consultas utilizam das coordenadas euclidianas em dados espaciais para verificar se o objeto de interesse está dentro da cobertura da rota. No entanto, essa abordagem não se aplica a dados puramente métricos, pois não se pode representá-los em um espaço n-dimensional. Por outro lado, uma função de distância pode inferir segmentos de retas sobre o espaço métrico, semelhante à associação de uma geometria. Neste contexto, o presente trabalho associa propriedades Euclidianas ao espaço métrico para definir a qualificação entre ponto e retas, de modo a expandir consultas contínuas por abrangência a este domínio de dados. Os experimentos evidenciam que esta associação permite consultas contínuas por abrangência eficientes em dados puramente métricos, suprimindo a lacuna das técnicas existentes na literatura.*

1. Introdução

Consultas que envolvem o conceito de trajetória são bastante investigadas na literatura, seja para fins de navegação ou para análise de deslocamento [Wang et al. 2021], usando dispositivos que agregam dados de Sistema de Posicionamento Global, sendo conhecidas como Consultas Contínuas [Xuan et al. 2008]. Consultas Contínuas são usadas apenas sobre dados multidimensionais, devido à natureza espacial de uma trajetória.

Entretanto, alguns tipos de dados, genericamente chamados de dados métricos, nem sempre possuem estrutura multidimensional (e.g. strings, proteínas), mas todos podem ser comparados por uma função de distância, que quantifica o quão dissimilar dois elementos são [Zezula et al. 2010]. Considerando que uma trajetória é definida como uma sequência de elementos que se conectam, partindo de um elemento até alcançar outro [Güting et al. 2025, Afonso et al. 2011], uma sequência de dados métricos estabelece o equivalente a uma trajetória. Por exemplo, uma frase pode ser entendida como uma trajetória através de uma sequência de palavras que a constitui. Analogamente, uma vacina de RNA mensageiro constitui um conjunto sequencial de proteínas.

Portanto, de forma análoga às consultas sobre dados espaciais, a hipótese deste trabalho é que Consultas Contínuas podem ser aplicadas a trajetórias de dados métricos. Por exemplo: *quais proteínas estão a até 1 limiar de alinhamento de uma sequência ordenada de proteínas (trajetória) que são sintetizadas por uma vacina de RNA mensageiro*?

Neste contexto, o objetivo deste trabalho é estender a execução de Consultas Contínuas por Abrangência para dados puramente métricos, isto é, que não podem ser representados no modelo de espaço *n*-dimensional. Especificamente, estas consultas recuperam dados métricos que estão a até um limiar máximo de uma trajetória formada por dados métricos. As principais contribuições deste trabalho são:

- A implementação de Consultas Contínuas por Abrangência em índices métricos.
- Definição da qualificação de nó em estruturas métricas usando o Teorema inverso do Teorema de Pitágoras para dividir o espaço em 3 partições que evita o uso de coordenadas para operar entre ponto e reta.
- Uso de dados métricos, como elementos de ponto e reta para esta Consulta.

Este artigo está dividido como segue: a Seção 2 apresenta os trabalhos correlatos. A Seção 3 apresenta a definição geométrica da proposta. A Seção 4 discute experimentos e resultados. A Seção 5 conclui o trabalho e delineia propostas futuras.

2. Referencial Teórico

Consultas Contínuas são usadas apenas sobre dados multidimensionais para recuperar objetos ao longo de uma trajetória, devido à natureza espacial dos dados. [Deng et al. 2011] classifica consultas em trajetórias que satisfaçam uma relação espaço-temporal entre trajetória e os seguintes elementos espaciais (ou vice-versa): por Pontos de Interesse (POI), por Região de Interesse (ROI) e por Trajetória de Interesse (TOI). A relação entre os objetos da consulta emprega critérios de vizinhança (*continuous k-nearest neighbour*: *k* pontos de interesse mais próximos a cada segmento) ou abrangência (*continuous range:* pontos a até uma distância máxima da trajetória) [Papadias et al. 2003, Xuan et al. 2008]. Alguns exemplos de consulta POI são: buscar os pontos dentro de um limiar de distância em um segmento de trajetória [Kalashnikov et al. 2002]; ou buscar o vizinho mais próximo de cada ponto em um segmento de trajetória [Tao et al. 2002, Chen et al. 2005]. Algumas variantes da consulta POI consideram fatores mais práticos no contexto de uma malha rodoviária, ou seja, a direção e o destino da viagem ao longo da trajetória [Chen et al. 2009, Shang et al. 2010].

O trabalho de [Wang et al. 2021] faz uma revisão sistemática sobre gerenciamento de dados de trajetórias, abrangendo contribuições em indexação com a *R-Tree* [Guttman 1984], decomposição do *pipeline* de processamento e uso de aprendizado profundo. Observa-se que as contribuições existentes na área envolvem o gerenciamento de dados de trajetória, exclusivamente em domínio espacial.

Em [Afonso et al. 2011] é apresentado o primeiro trabalho a mostrar a consulta k-vizinho mais próximo de uma trajetória indexando a *M-Tree* [Ciaccia et al. 1997] com a distância de edição com penalidade real. No entanto, seus experimentos utilizaram dados espaciais. Outro exemplo de uso de dados espaciais com função de distância de edição é [Chen et al. 2005], onde as trajetórias foram obtidas pelo rastreamento das pontas dos dedos das formando palavras da língua de sinais. [Güting et al. 2025] propõe uma nova estrutura de indexação métrica em memória, a *N-Tree* que particiona o espaço métrico usando múltiplos centros em dados usando uma nova função de distância métrica para trajetórias de objetos móveis. Os experimentos usam dados espaciais para realizar consulta contínua por abrangência e aos k-vizinhos mais próximos. Embora estruturas de índices sejam empregadas como forma de armazenar informações e acelerar o desempenho das consultas, algumas propostas realizam operações de podas no processamento da consulta, envolvendo a divisão do espaço de busca em duas regiões [Xuan et al. 2008].

Este trabalho investigou o tipo de consulta por pontos de interesse em trajetória com limiar de cobertura. Entretanto, pela análise das distâncias entre uma trajetória e os pontos de interesse, optou-se por combinar métodos de acesso métricos com as propriedades Euclidianas para dividir o espaço em três regiões, pela aplicação do Teorema inverso do Teorema de Pitágoras, para contornar a ausência de dimensionalidade dos dados.

3. Consultas Contínuas por Abrangência em Dados Métricos

Inicialmente, a primeira condição para utilização da proposta é a definição da qualificação do nó em estruturas métricas operando sobre pontos de interesse e segmentos de retas. Os pontos de interesse são os elementos do conjunto, que são indexados; e as retas são inferidas e formadas por segmentos da trajetória, definidos por dois elementos consecutivos.

$$a = dist(x, y) \quad (1)$$

$$b = dist(x, u) \quad (2)$$

$$c = dist(y, u) \quad (3)$$

$$h = (2A)/2 \quad (4)$$

$$A = \sqrt{P(P-a)(P-b)(P-c)} \quad (5)$$

$$P = (a + b + c)/2 \quad (6)$$

$$D1 = \alpha \geq \pi/2 \leftrightarrow c^2 \geq a^2 + b^2 \quad (7)$$

$$D3 = \beta \geq \pi/2 \leftrightarrow b^2 \geq a^2 + c^2 \quad (8)$$

$$D2 = \neg D1 \ \wedge \ \neg D3 \quad (9)$$

$$\{u \mid u \in U \wedge \ \forall \langle x, y \rangle \ \in T \wedge ((D1 \wedge b \leq \varepsilon) \vee (D3 \wedge c \leq \varepsilon) \vee (D2 \wedge h \leq \varepsilon))\} \quad (10)$$

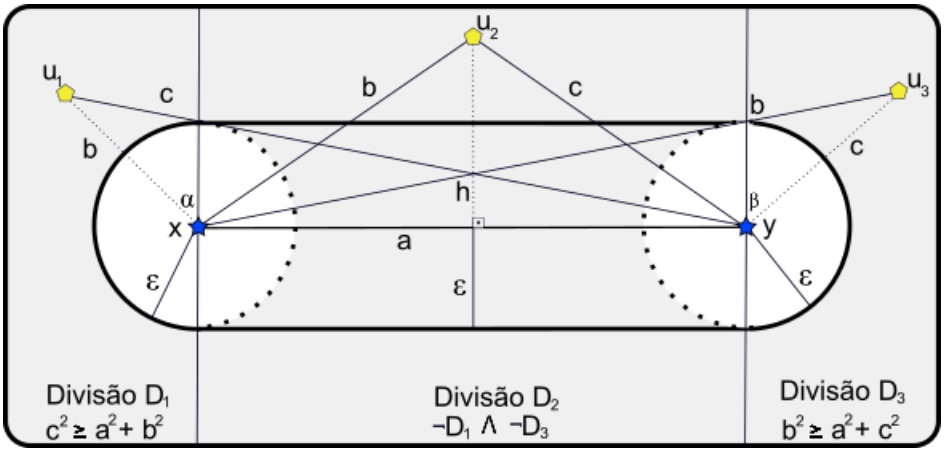

Figura 1. Qualificação na Consulta Contínua Por Abrangência.

Navegando do segmento da trajetória x para y da Figura 1, tem-se a definição de um limiar de abrangência ε que especifica uma região de cobertura máxima na qual serão identificados pontos de interesse ao longo da trajetória. Os pontos de interesse podem estar localizados em uma de 3 possíveis divisões: D1, onde são encontrados pontos que antecedem x; D3, pontos que sucedem y; D2, pontos entre x e y.

A divisão D1 contém pontos de interesse que estão à esquerda da perpendicular originada pelo ponto inicial *x*, estabelecida na equação (7); a divisão D3 contém pontos de interesse que estão à direita da perpendicular originada pelo ponto final *y*, estabelecida na equação (8); enquanto que na divisão D2 estão os pontos que não estão na divisão D1 e D3, estabelecida na equação (9). Para identificar a divisão onde o ponto se encontra, pode-se verificar os ângulos da origem *x* e do final *y* do segmento da trajetória que são gerados pela triangulação com o ponto de interesse *u*.

Como pode ser visto na Figura 1, para a situação D1 o ângulo α é obtuso (maior que 90 graus) para a reta oposta *c* formada pela distância entre os pontos *u* e *y* (Equação (3)). Enquanto que em D3 o ângulo β é obtuso para a reta oposta *b* formada pela distância entre *u* e *x* (Equação (2)). Para simplificar, é possível determinar se um triângulo é obtuso, sem calcular valor angular, usando Teorema inverso do Teorema de Pitágoras que pode ser visto nas Equações (7) e (8).

Para existir qualificação (Algoritmo 1): na divisão D1, a reta *b* da Equação (2) deve ser menor ou igual a ε; na divisão D3, a reta *c* da Equação (3) deve ser menor ou igual a ε; e na divisão D2 deve-se calcular a perpendicular *h* através da fórmula de Heron das Equações (4, 5 e 6). A partir dessas definições, é estabelecido o cálculo relacional para a Consulta contínua por Abrangência em uma trajetória através da Equação (10).

Algoritmo 1: Qualifica(ponto, reta, cov, ε) :
a=dist(reta.inicio, reta.fim); b=dist(reta.inicio, u); c=dist(reta.fim,u);
se ($c^2 \geq a^2 + b^2$) **então return** ($b \leq cov + \varepsilon$);
senão se ($b^2 \geq a^2 + c^2$) **então return** ($c \leq cov + \varepsilon$);
senão P=(a+b+c)/2; A=$\sqrt{P(P-a)(P-b)(P-c)}$; $h = (2A)/2$; **return** ($h \leq cov + \varepsilon$);

O Algoritmo 2 apresenta a implementação da Consulta Contínua por Abrangência que percorre blocos índices e folhas de uma estrutura M-Tree ou Slim-Tree [Traina Jr. et al. 2002] para verificar qualificações de pontos que estão até um limiar ε da trajetória.

Algoritmo 2: CCPA-Metric(Bloco, Trajetória, ε, resposta) :
```
para cada entrada(e) ∈ bloco faça
  para cada segmento(r) ∈ Trajetória faça
    se bloco é índice então
      se Qualifica(e.ponto, r, e.cobertura, ε) então
        CCPA-Metric(e.subBloco, Trajetória, ε, resposta);
        interrompa loop Trajetória;
    senão
      se Qualifica(e.ponto, r, 0, ε) então
        adicione a resposta ⟨e.ponto, r⟩;
        interrompa loop Trajetória;
```

4. Resultados

Esta seção avalia a proposta em relação ao tempo de execução e quantidade de cálculos de distância. Os testes foram executados sobre o Windows 11, Intel Core i7-13700, 16 GB RAM, 250 GB NVME, e os algoritmos implementados em Java. Os conjuntos de dados são: CodePoint [CodePoint 2025] de natureza espacial contendo 1666774 Latitudes e Longitudes; LibreOffice [LibreOffice 2025] de natureza métrica contendo 318643 palavras; e UniProt [UniProtKB 2025] que foi restrito às proteínas com tamanho máximo de 46 aminoácidos, resultando 100000 elementos. Para o experimento sobre o conjunto CodePoint foram geradas 10 séries sintéticas de trajetórias variando a quantidade de pontos em cada série, de 300 até 3000 pontos na rota, em incrementos de 300. A geração de rotas realiza uma consulta 100-NN sobre um ponto de interesse aleatório e, a seguir, seleciona-se aleatoriamente um ponto sucessor da trajetória. O processo se repete até atingir o total de pontos da série. Para cada série, são geradas 50 trajetórias para cômputo da média das medições. O conjunto CodePoint foi indexado, separadamente, nas estruturas R-Tree, M-Tree e Slim-Tree, com blocos de 1 KB e métrica L_2. A Figura 2(a) apresenta o tempo de execução da Consultas Contínuas por Abrangência. A Slim-Tree foi, respectivamente, 74% e 100% mais rápida que as R-Tree e M-Tree. A Figura 2(b) mostra o total de verificações, sendo que a R-Tree realizou, respectivamente, 139% e 10% menos verificações que as M e Slim Trees, porém, com maior custo computacional.

Para o conjunto LibreOffice, as trajetórias foram definidas a partir das frases da obra "Memórias Póstumas de Brás Cubas". As trajetórias, formadas por frases, foram divididas em 10 séries, cada uma contendo 2, 4, 8, 12, 16, 24, 32, 40, 52 e 151 palavras. Para cada série, existe uma média de 300 trajetórias, que são usadas no cômputo da média. O conjunto LibreOffice foi indexado, separadamente, nas estruturas M-Tree e Slim-Tree, com blocos de 8 KB e métrica L_{Edit}. A Figura 2(c) mostra o tempo de execução das consultas, sendo que a Slim-Tree foi 8,5% mais rápida M-Tree. A Figura 2(d) mostra que a Slim-Tree realizou 7,7% menos cálculos de distância que a M-Tree.

Para realizar o experimento sobre o conjunto UniProt foram geradas 10 séries sintéticas de trajetórias usando o mesmo procedimento do conjunto CodePoint, mas variando de 10 até 100 proteínas na rota, em incrementos de 10. Para cada série, são geradas 50 trajetórias para cômputo da média das medições. O conjunto UniProt foi indexado, separadamente, nas estruturas M-Tree e Slim-Tree, com blocos de 8 KB e métrica L_{Edit} ponderada com mPAM [Xu e Miranker 2004] que avalia semanticamente a evolução da substituição de um aminoácido por outro, usando um modelo Markoviano. A Figura 2(e) apresenta o tempo de execução da Consultas, sendo que a Slim-Tree foi 31% mais rápida M-Tree. A a Slim-Tree realizou 24% menos cálculos de distância que a M-Tree (Figura 2(f)).

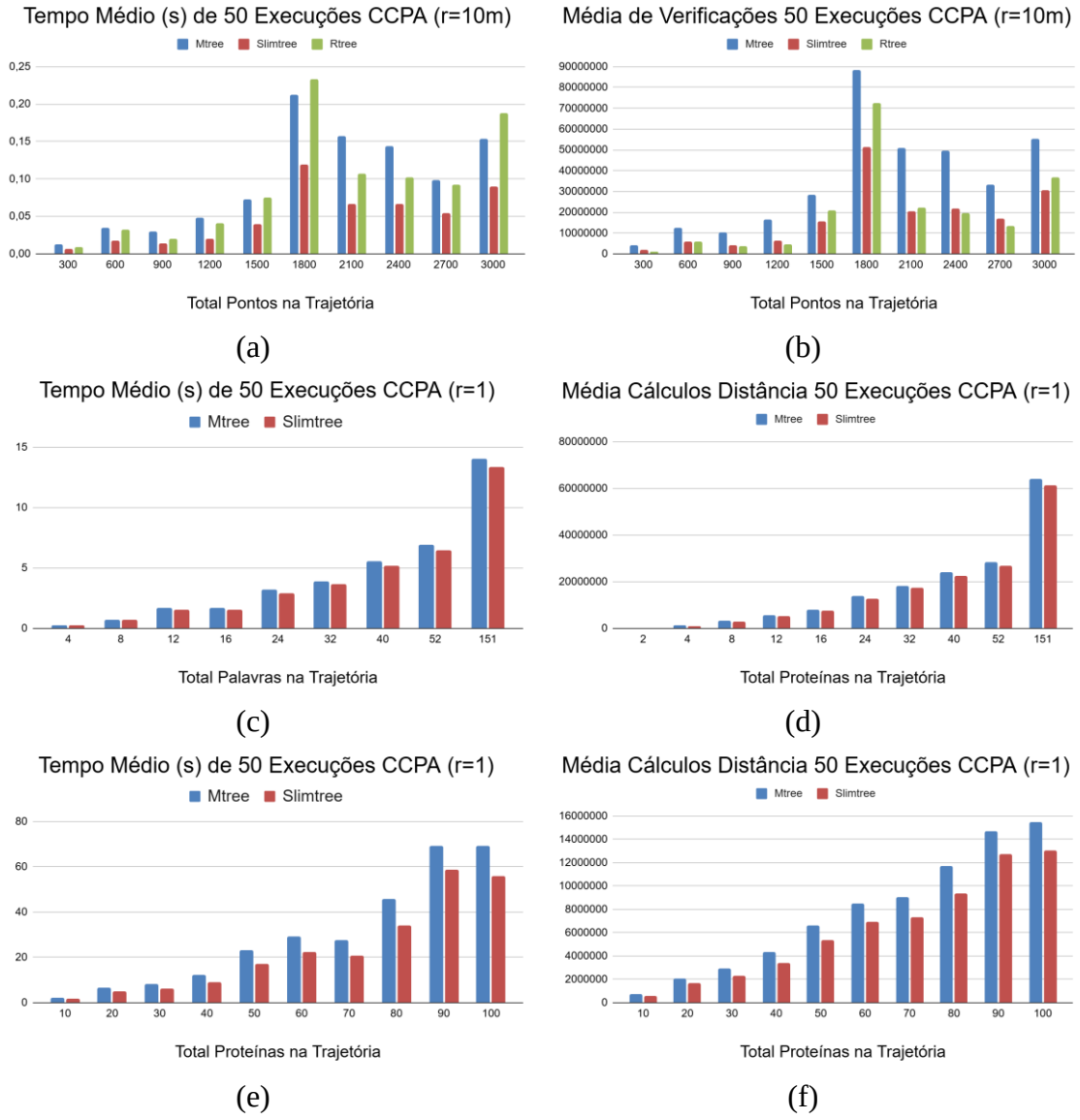

Figura 2. Gráficos da Média dos Tempos e das Verificações da CCPA.

5. Conclusão

Este trabalho estendeu a consulta contínua por abrangência para dados métricos usando o Teorema inverso do Teorema de Pitágoras para a implementação nas estruturas métricas. Foram realizados três experimentos: com rotas espaciais sintéticas usando a métrica L_2 para validar a corretitude da proposta; com rotas reais formadas por frases usando a métrica L_{Edit}; com rotas sintéticas formadas por sequências de proteínas usando métrica L_{Edit} ponderada semanticamente pela evolução de aminoácidos. Os resultados mostram que a Slim-Tree obteve desempenho que a melhor estrutura em relação ao tempo 74% na base de pontos; 8,5% na base de palavras; e 31% na base de proteínas. Como trabalhos futuros, observa-se que a presente proposta abre novas possibilidades de investigação de consultas contínuas, tais como: identificação dos pontos de interesse mais próximos de cada segmento de trajetória; estender a investigação para os tipos ROI e TOI; uso Unidade de Processamento Gráfico para aceleração da Consulta Contínua.

References

Afonso, F., Barbosa, F., and Rodrigues, A. (2011). Trajectory data similarity with metric data structures. In Geographical Inf. Science Research United Kingdom, 9p.

Ciaccia, P., Patella, M., and Zezula, P. (1997). M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In Proc. Int. Conf. VLDB, pages 426–435, Morgan Kaufmann, San Francisco, CA, USA.

Chen, L., Özsu, M. T., and Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In Proc. ACM SIGMOD, pages 491–502, New York, NY, USA.

Chen, Z., Shen, H. T., Zhou, X., and Yu, J. X. (2009). Monitoring path nearest neighbor in road networks. In Proc. ACM SIGMOD, pages 591–602, New York, NY, USA.

CodePoint, Open CSV. (2025). https://osdatahub.os.uk/downloads/open/CodePointOpen

Deng, K., Xie, K., Zheng, K., and Zhou, X. (2011). Trajectory Indexing and Retrieval, pages 35–60. Springer, New York, NY.

Guttman. A. (1984). R-trees: a dynamic index structure for spatial searching. SIGMOD Rec., 14 (2).

Güting, R. H., Das, S. K., Valdés, F., and Ray, S. (2025). Exact trajectory similarity search with n-tree: An efficient metric index for knn and range queries. ACM Trans. Spatial Algorithms Syst., 11(1).

Kalashnikov, D., Prabhakar, S., Hambrusch, S., and Aref, W. (2002). Efficient evaluation of continuous range queries on moving objects. In Proc. DEXA, Berlin, Springer.

LibreOffice. (2025). https://github.com/LibreOffice/dictionaries

Papadias, D., Zhang, J., Mamoulis, N., and Tao, Y. (2003). Query processing in spatial network databases. In VLDB Conf., pages 802–813. Morgan Kaufmann, San Francisco.

Shang, S., Deng, K., and Xie, K. (2010). Best point detour query in road networks. In Proc. 18th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems, pages 71–80, New York, NY, USA.

Tao, Y., Papadias, D., and Shen, Q. (2002). Continuous nearest neighbor search. In Proc. 28th Int. Conf. on VLDB, pages 287–298. Morgan Kaufmann, San Francisco.

Traina Jr., C., Traina, A. J. M., Faloutsos, C., and Seeger, B. (2002). Fast Indexing and Visualization of Metric Data Sets using Slim-Trees. IEEE TKDE. 14(2).

UniProtKB, TrEMBL Fasta. (2025). https://www.uniprot.org/help/downloads

Wang, S., Bao, Z., Culpepper, J. S., and Cong, G. (2021). A survey on trajectory data management, analytics, and learning. ACM Comput. Surv., 54(2).

Xu, W., and Miranker, D. P. (2004). A metric model of amino acid substitution. Bioinformatics. UK, 20(8).

Xuan, K., Zhao, G., Taniar, D., and Srinivasan, B. (2008). Continuous range search query processing in mobile navigation. In Int. Conf. on Parallel and Dist. Syst, pages 361–368.

Zezula, P., Amato, G., Dohnal, V., and Batko, M. (2010). Similarity Search: The Metric Space Approach. Springer Publishing Company, Incorporated, 1st edition.