Informe 1 2025

Detección de Heavy Hitters usando Sketches

Ignacio Barría Concha - Nicolás Jarpa Jeldres - Nicolás Pino Leal

Los archivos asociados a este trabajo con el objetivo de identificar kmers canonicos como heavy hitters usando sketches estan en el siguiente enlace: https://github.com/Pinox084/Tarea1TMGVD

Actividad 1: Preparación de datos y ground truth

Objetivo

El objetivo de esta actividad fue definir el conjunto de *k-mers*, calcular un *ground truth* de frecuencias exactas en un subconjunto de datos representativo y establecer umbrales de *heavy hitters (HH)*.

Metodología

1. Procesamiento de genomas

- Se trabajó con archivos de genomas en formato **FASTA**.
- Se limpiaron las secuencias dejando únicamente bases **A, C, G, T**, descartando aquellas posiciones con caracteres ambiguos como **N**.
- Cada archivo listado en listaarchivos.txt fue procesado de manera independiente.
- Conservábamos un 1MB de memoria para HH de cada archivo FASTA y los juntamos en un solo archivo.

2. Extracción de k-mers canónicos

- Se implemento un archivo principal llamado kmersheader.hpp para almacenar funciones que permiten codificar los kmers en base 2 y obtener sus formas canónicas.
- La función que entrega los kmers canónicos recibe una sola una línea string de tamaño k que compara el kmer y su complemento y retornar su forma canónica.

- Se consideró tanto el *k-mer* como su reverso complementario, guardando siempre el **canónico** (el menor en orden lexicográfico).

3. Cálculo de frecuencias exactas

- Se desarrolló el código que realiza un **conteo exacto de k-mers** utilizando un unordered_map<u64,uint64_t>. Donde estas, se obtienen sus frecuencias exactas, del archivo GTExacto.cpp.
- El resultado es un archivo con todas las frecuencias exactas (out_exact.txt), donde cada línea contiene:

KMER	Frecuencia
…	…
11223002121021211011 2	10
…	…

Guardados en formato de 2 bits

4. Definición de umbral ϕ y obtención de heavy hitters

- El umbral se definió como:

$$f(x) \geq \varphi N$$

donde f(x) es la frecuencia de un *k-mer* y N el total de *k-mers* válidos procesados en el archivo.

- Se probaron valores de φ, y coincidimos en el φ fijo de 2e-6, esto gracias a ensayo y error en los algoritmos de countSketch y towerSketch con CountMin y Conservative Update (TS CMCU). Esto debido a que en los dos scripts de la actividad 2 el phi afectaba mucho más de lo que afectaba en la actividad 1, por tanto, llegamos a una conclusión de que teniendo un φ=2e-6 dejaría buenos resultados para todo lo pedido.
- El conjunto de heavy hitters de cada archivo se guardó en HHexactos.txt y HHexactos21.txt, respectivamente para un K=31 y un K=21.
- Para cumplir con la restricción de memoria, se implementó un límite de 1 MB por archivo, y se junta en los txt resultantes HHexactos.txt y HHexactos21.txt.

Resultados

1. **Frecuencias exactas (Ground Truth)**

- Se obtuvo para cada archivo el conteo exacto de todos los *k-mers* .
- Esto constituye el **ground truth** que servirá para evaluar la precisión de los algoritmos de sketch.

2. **Heavy Hitters**

- Para cada archivo se generó la lista de HH exactos con su frecuencia.
- Los resultados se almacenaron en un único archivo de salida HHexactos.txt (para k=31) y HHexactos21.txt (para k=21), con secciones separadas por archivo.
- Cada bloque contiene como máximo 1 MB de información para cumplir con las restricciones de la actividad.

3. **Medición de memoria**

- Se registró la memoria aproximada utilizada por el unordered_map durante el conteo exacto.

Actividad 2

Se implementaron las estructuras CS y TS-CMCU con funciones insert(kmer) y estimate(kmer) considerando múltiples filas por tamaño de contadores, a continuación, se presenta el seudocódigo de ambas implementaciones.

Pseudocódigo – Count Sketch

Estructura CountSketch(d, w):

tabla[d][w] ← 0

hashSeeds[j] ← valores distintos para cada fila j

signSeeds[j] ← valores distintos para cada fila j

Función SIGN(key, j):

h ← Hash(key, signSeeds[j])

Si (h es impar) retornar +1

Si (h es par) retornar -1

Procedimiento UPDATE(key, delta):

Para cada fila j = 0 ... d-1:

h ← Hash(key, hashSeeds[j])

col ← h mod w

s ← SIGN(key, j)

tabla[j][col] ← tabla[j][col] + delta * s

Función ESTIMATE(key):

Para cada fila j = 0 ... d-1:

h ← Hash(key, hashSeeds[j])

col ← h mod w

s ← SIGN(key, j)

est[j] ← tabla[j][col] * s

retornar mediana(est[0..d-1])

update(key, delta) → Inserción / actualización

Para cada una de las “d” filas de la tabla:

1. Se calcula una posición hash col para la clave.
2. Se obtiene un signo aleatorio determinístico +1 o -1 usando otra función hash.
3. Se actualiza el contador de la celda correspondiente con delta * signo.

Esto permite compensar colisiones: si dos claves distintas caen en la misma celda, sus contribuciones pueden tener signos distintos, reduciendo el sesgo.

estimate(key) → Estimación de frecuencia

Para cada fila:

1. Se calcula la misma celda y el mismo signo que se usó en la inserción.
2. Se recupera el valor almacenado y se multiplica por el signo → esto "deshace" el efecto del signo aplicado en update.
3. Se toma la mediana de las d estimaciones → esto reduce el efecto de posibles colisiones grandes en alguna fila.

El resultado es una estimación no sesgada de la frecuencia de la clave, con alta probabilidad de estar cerca del valor real si "d" y "w" son lo suficientemente grandes.

Pseudocódigo -TS-CMCU

```
Estructura CountMinCU(width, rows):
    tabla[rows][width] ← 0
    seeds[r] ← valores distintos para cada fila r

Procedimiento INSERT(key):
    Para cada fila r = 0 ... rows-1:
        h ← Hash(key, seeds[r])
        idx[r] ← h mod width
    minv ← ∞
    Para cada fila r:
        minv ← min(minv, tabla[r][idx[r]])
    Para cada fila r:
        Si tabla[r][idx[r]] == minv:
            tabla[r][idx[r]] ← tabla[r][idx[r]] + 1

Función ESTIMATE(key):
    ans ← ∞
    Para cada fila r:
        h ← Hash(key, seeds[r])
```

idx ← h mod width

ans ← min(ans, tabla[r][idx])

Si ans == ∞ retornar 0

retornar ans

insert(key) → Inserción con Conservative Update

1. Se calcula en qué posición (idx) cae la clave en cada una de las rows filas.
2. Se busca el **mínimo valor actual** entre esas posiciones.
3. Se incrementan **solo las celdas que tienen el valor mínimo**.

Esto evita **sobreestimaciones innecesarias**: en el Count-Min clásico se incrementan todas las celdas, lo que puede amplificar errores por colisiones. Con Conservative Update solo se incrementan las celdas más "rezagadas", manteniendo las otras igual.

estimate(key) → Estimación

1. Se calcula la celda correspondiente en cada fila.
2. Se devuelve el **mínimo de todos los contadores** asociados a la clave.

El mínimo actúa como una cota superior del conteo real, ya que por colisiones el valor en cada celda solo puede **sobreestimar**, nunca subestimar. Tomar el mínimo entre varias filas reduce el error.

Estructura TowerSketch(widths[], rows_per_level[], seed0):

Para cada nivel i:

Crear CountMinCU con width = widths[i], rows = rows_per_level[i], seed = seed0 + i

Añadirlo a levels

Procedimiento INSERT(key):

Para cada nivel en levels:

nivel.INSERT(key)

Función ESTIMATE(key):

best ← ∞

Para cada nivel en levels:

e ← nivel.ESTIMATE(key)

best ← min(best, e)

Si best == ∞ retornar 0

retornar best

Tower Sketch combina varios Count-Min Sketch con distintos anchos y números de filas (niveles).

- Cada inserción se propaga por todos los niveles.
- Para estimar, se toma el mínimo entre las estimaciones de todos los niveles.

Esto mejora la precisión y estabilidad respecto a usar un único sketch, especialmente en rangos amplios de frecuencias (muy útiles para heavy hitters o k-mers con distribuciones sesgadas).

Calibraciones

Count Sketch:
Se calibra probando distintas combinaciones de número de filas (`d`) y ancho de tabla (`w`). Para cada configuración, se insertan todos los k-mers de un grupo seleccionado de archivos y se comparan las estimaciones con las frecuencias reales, midiendo error (MAE/MRE) y calidad en la detección de heavy hitters (Precisión, Recall, F1). Los parametros utilizados son los siguientes:

vector<int> ds = {9, 10, 11, 15, 20};

vector<int> ws = {10000, 20000, 50000, 100000, 200000, 500000, 1000000, 2000000};

A continuación, se muestra “d” y “w” de prueba para un φ de 2e-06 de tamaños del k-mer de 21.

N°	d	w	Tamaño (bytes)	MAE	MRE	Precisión	Recall	F1
1	9	1 000 000	72 000 000	0.254749	0.254163	0.0807175	1.000	0.149378
2	15	500 000	60 000 000	0.350795	0.350006	0.0380952	0.667	0.0720721
3	20	2 000 000	320 000 000	0.0225919	0.0225223	0.18	1.000	0.305085
4	10	100 000	8 000 000	1.24864	1.24586	8.51396e-05	0.667	0.0001703
5	11	500 000	44 000 000	0.432045	0.431076	0.0241636	0.722	0.0467626
6	9	200 000	14 400 000	0.858456	0.856542	0.000476304	0.556	0.0009518
7	20	1 000 000	160 000 000	0.107176	0.106925	0.130435	1.000	0.230769
8	15	1 000 000	120 000 000	0.147619	0.147280	0.112583	0.944	0.201183
9	10	500 000	40 000 000	0.490636	0.489526	0.0094518	0.833	0.0186916
10	9	50 000	3 600 000	1.78661	1.78267	2.92946e-05	0.333	5.8584e-05

A continuación, se muestra “d” y “w” de prueba para un φ de 2e-06 de tamaños del k-mer de 31.

N°	d	w	Tamaño (bytes)	MAE	MRE	Precisión	Recall	F1
1	20	500 000	80 000 000	0.240139	0.239809	0.00348172	0.875	0.00693584
2	9	500 000	36 000 000	0.423760	0.423208	0.0009612 54	0.8125	0.00192024
3	11	200 000	17 600 000	0.694683	0.693736	0.0002057 76	0.875	0.00041146
4	15	20 000	2 400 000	2.017230	2.014600	0.0000255 86	0.5625	0.00005117
5	9	1 000 000	72 000 000	0.199586	0.199306	0.0055086 30	0.9375	0.01095290
6	11	10 000	880 000	3.314140	3.309930	0.0000198 89	0.5625	0.00003978
7	20	1 000 000	160 000 000	0.071852	0.071729	0.0115108 00	1.0000	0.02275960
8	15	500 000	60 000 000	0.294462	0.294074	0.0033975 10	0.9375	0.00677048
9	10	10 000	800 000	3.652150	3.647460	0.0000162 74	0.625	0.00003255
10	11	1 000 000	88 000 000	0.158730	0.158496	0.0078125 00	1.0000	0.01550390

Tower Sketch:

Se calibra definiendo varios niveles con anchos decrecientes y filas fijas, ajustando los parámetros según ϕ, δ y la memoria disponible. Se evalúa igual que Count Sketch, tomando métricas de error y heavy hitters para comparar configuraciones. Los parámetros utilizados fueron los siguientes

vector<vector<size_t>> configs_widths = {

{1<<20, 1<<19, 1<<18, 1<<17, 1<<16},

{1<<21, 1<<20, 1<<19, 1<<18, 1<<17},

{1<<23, 1<<22, 1<<21, 1<<20, 1<<19},

{1<<24, 1<<23, 1<<22, 1<<21, 1<<20, 1<<19, 1<<18}

}

vector<vector<size_t>> configs_rows = {

{5, 5, 4, 4, 3},

{6, 5, 5, 4, 4},

{7, 6, 6, 5, 5},

{10, 8, 7, 7, 6, 6, 5}

}

A continuación, se muestra d y w de prueba para un φ de 2e-06 de tamaños del k-mer de 21.

Config	N	MAE	RMSE	Max Abs Error	Max Rel Error	Precision	Recall	F1Score
cfg 1	1302461	0.00409202	0.0742833	12	12	0.323442	1	0.488789
cfg 3	1514500	0.013299	0.121375	6	6	0.188119	1	0.316667
cfg 2	3982388	0.0185966	0.209265	35	35	0.390465	1	0.561633

cfg 0	2277733	0.174 591	0.454 005	29	29	0.460581	1	0.630682
cfg 1	1942545	0.009 11808	0.251 977	41	41	0.240964	1	0.38835
cfg 2	1580689	0.002 05053	0.047 1459	7	7	0.638889	1	0.779661
cfg 1	3980036	0.100 015	0.372 99	76	76	0.0797721	1	0.147757
cfg 2	2284665	0.006 94263	0.098 1644	12	12	0.0740741	1	0.137931
cfg 0	1514500	0.053 1368	0.234 341	6	6	0.180952	1	0.306452
cfg 3	3982388	0.018 5958	0.209 263	35	35	0.390465	1	0.561633

A continuación, se muestra d y w de prueba para un φ de 2e-06 de tamaños del k-mer de 31.

Config	**N**	**MAE**	**RMSE**	**Max Abs Error**	**Max Rel Error**	**Precision**	**Recall**	**F1Score**
cfg 2	1260 958	0.012 9983	0.1198 78	7	7	0.0264026	1	0.0514469
cfg 1	1084 253	0.003 7918	0.0759 957	9	9	0.186047	1	0.313725
cfg 3	1316 363	0.002 0622 7	0.0472 008	5	5	0.121212	1	0.216216
cfg 0	3318 223	0.454 088	0.7172 79	29	29	0.193969	1	0.324915
cfg 1	1897 122	0.018 3609	0.2808 1	25	25	0.209479	1	0.346395

cfg 3	1903 784	0.006 1672 5	0.0950 366	11	11	0.278607	1	0.435798
cfg 0	1618 601	0.063 3776	0.4062 76	35	35	0.0786651	1	0.145856
cfg 0	7978 64	0.002 8542 6	0.0545 628	2	2	0.0462527	1	0.0884159
cfg 1	3316 461	0.058 5379	0.3212 62	52	52	0.0018832 4	1	0.0037594
cfg 2	3318 223	0.018 6584	0.2329 47	29	29	0.199664	1	0.332867

Actividad 3: Detección de Heavy Hitters y Validación

Se ejecutan las dos estructuras usando los parametros con mejor rendimiento que se obtuvieron tras la calibración, para la detección de heavy hitters se usaron los siguientes parametros:

- Φ = 2e-6 (0.000002)
- Count Sketch: d= 11, witdth = 2.000.000
- TS-CMCU: widths = {1<<22, 1<<21, 1<<20, 1<<19, 1<<18}, rows = {6, 6, 5, 5, 4}

TS-CMCU con k = 21

idx	name	TP	FP	FN	precision	recall	f1	hh_exact	hh_est
1	GCA_006152045.1_ASM615204v1_genomic.fna	1	524	0	0,001905	1	0,003802	1	525
2	GCA_018421455.1_ASM1842145v1_genomic.fna	238	954	0	0,199664	1	0,332867	238	1192
3	GCA_020118255.1_ASM2011825v1_genomic.fna	56	1172	0	0,045603	1	0,087227	56	1228
4	GCA_021919605.1_PDT001092240.1_genomic.fna	42	916	0	0,043841	1	0,084	42	958
5	GCA_021953145.1_PDT001020446.1_genomic.fna	17	746	0	0,02228	1	0,04359	17	763
6	GCA_021972575.1_PDT001013406.1_genomic.fna	14	795	0	0,017305	1	0,034022	14	809
7	GCA_022062785.1_PDT000876120.1_genomic.fna	30	776	0	0,037221	1	0,07177	30	806

8	GCA_023315275.2_PDT001299634.2_genomic.fna	17	904	0	0,018458	1	0,036247	17	921
9	GCA_024452925.1_PDT001370221.1_genomic.fna	55	927	0	0,056008	1	0,106075	55	982
10	GCA_024732165.1_PDT001378927.1_genomic.fna	44	825	0	0,050633	1	0,096386	44	869
11	GCA_026006075.1_ASM2600607v1_genomic.fna	495	1004	0	0,33022	1	0,496489	495	1499
12	GCA_026006095.1_ASM2600609v1_genomic.fna	533	954	0	0,35844	1	0,527723	533	1487
13	GCA_026305215.1_PDT001493239.1_genomic.fna	1	555	0	0,001799	1	0,003591	1	556
14	GCA_031045075.2_PDT001744190.2_genomic.fna	53	910	0	0,055036	1	0,104331	53	963
15	GCA_032567175.1_ASM3256717v1_genomic.fna	48	855	0	0,053156	1	0,100946	48	903
16	GCA_033106465.1_PDT001960943.1_genomic.fna	56	940	0	0,056225	1	0,106464	56	996
17	GCA_037052065.1_ASM3705206v1_genomic.fna	0	411	0	0	0	0	0	411
18	GCA_037203645.1_ASM3720364v1_genomic.fna	111	943	0	0,105313	1	0,190558	111	1054
19	GCA_043678295.1_ASM4367829v1_genomic.fna	80	937	0	0,078663	1	0,145852	80	1017
20	GCA_043678335.1_ASM4367833v1_genomic.fna	306	980	0	0,237947	1	0,384422	306	1286
21	GCA_043678355.1_ASM4367835v1_genomic.fna	603	895	0	0,402537	1	0,574012	603	1498
22	GCA_043678375.1_ASM4367837v1_genomic.fna	483	1000	0	0,325691	1	0,491353	483	1483
23	GCA_043950085.1_ASM4395008v1_genomic.fna	20	555	0	0,034783	1	0,067227	20	575
24	GCA_943323015.2_HRS-ES2-bin-440_genomic.fna	576	3454	0	0,142928	1	0,250109	576	4030
25	GCA_943323415.2_HRS-ES4-bin-10_genomic.fna	2366	3622	0	0,395124	1	0,566435	2366	5988
26	GCA_943323865.2_HRS-ES8-bin-118_genomic.fna	504	3380	0	0,129763	1	0,229717	504	3884
27	GCA_943325355.2_HRSG-E10-bin-12_genomic.fna	8	656	0	0,012048	1	0,02381	8	664
28	GCA_943326675.2_HRSG-E12-bin-108_genomic.fna	71	664	0	0,096599	1	0,176179	71	735
29	GCA_943327095.2_HRSG-E11-bin-71_genomic.fna	12	403	0	0,028916	1	0,056206	12	415
30	GCA_943327695.2_HRS-HS3-bin-186_genomic.fna	1	106	0	0,009346	1	0,018519	1	107

31	GCA_943329525.2_HRS-HS6-bin-101_genomic.fna	108	1131	0	0,087167	1	0,160356	108	1239
32	GCA_943329805.2_HRSG-H13-bin-193_genomic.fna	442	1668	0	0,209479	1	0,346395	442	2110
33	GCA_943331045.1_HRSG-E2-bin-118_genomic.fna	16	116	0	0,121212	1	0,216216	16	132
34	GCA_943331235.2_HRSG-H12-bin-94_genomic.fna	24	496	0	0,046154	1	0,088235	24	520
35	GCA_943332185.1_HRS-HS9-bin-103_genomic.fna	67	1210	0	0,052467	1	0,099702	67	1277
36	GCA_943332625.1_HRS-HS9-bin-149_genomic.fna	11	157	0	0,065476	1	0,122905	11	168
37	GCA_943333245.1_HRSG-E1-bin-8_genomic.fna	1673	8581	0	0,163156	1	0,28054	1673	10254
38	GCA_943333305.2_HRSG-H6-bin-46_genomic.fna	26	2613	0	0,009852	1	0,019512	26	2639
39	GCA_943334125.2_HRSG-E6-bin-23_genomic.fna	144	1602	0	0,082474	1	0,152381	144	1746
40	GCA_943334475.2_HRSG-H3-bin-66_genomic.fna	188	4602	0	0,039248	1	0,075532	188	4790
41	GCA_943334715.2_HRSG-H6-bin-64_genomic.fna	56	145	0	0,278607	1	0,435798	56	201
42	GCA_943334795.2_HRSG-H6-bin-28_genomic.fna	48	210	0	0,186047	1	0,313725	48	258
43	GCA_945860935.1_GE-03apr19-182_genomic.fna	0	344	0	0	0	0	0	344
44	GCA_945867275.1_TH-11nov19-195_genomic.fna	41	699	0	0,055405	1	0,104994	41	740
45	GCA_945870865.1_TrH-03may19-115_genomic.fna	108	1014	0	0,096257	1	0,17561	108	1122
46	GCA_945870955.1_MoE-23oct19-325_genomic.fna	66	754	0	0,080488	1	0,148984	66	820
47	GCA_945872895.1_MaE-04nov19-161_genomic.fna	16	590	0	0,026403	1	0,051447	16	606
48	GCA_945874295.1_AH-24oct19-10_genomic.fna	25	150	0	0,142857	1	0,25	25	175
49	GCA_945877175.1_ZE-13nov19-95_genomic.fna	174	3410	0	0,048549	1	0,092602	174	3584
50	GCA_945902215.1_MoH-23oct19-118_genomic.fna	205	1472	0	0,122242	1	0,217853	205	1677

Count Sketch con k = 21

idx	name	TP	FP	FN	precision	recall	f1	hh_exact	hh_est
1	GCA_006152045.1_ASM615204v1_genomic.fna	1	515	0	0,001938	1	0,003868	1	516
2	GCA_018421455.1_ASM1842145v1_genomic.fna	231	973	7	0,19186	0,970588	0,320388	238	1204
3	GCA_020118255.1_ASM2011825v1_genomic.fna	53	1190	3	0,042639	0,946429	0,081601	56	1243
4	GCA_021919605.1_PDT001092240.1_genomic.fna	42	925	0	0,043433	1	0,083251	42	967
5	GCA_021953145.1_PDT001020446.1_genomic.fna	17	726	0	0,02288	1	0,044737	17	743
6	GCA_021972575.1_PDT001013406.1_genomic.fna	13	762	1	0,016774	0,928571	0,032953	14	775
7	GCA_022062785.1_PDT000876120.1_genomic.fna	29	780	1	0,035847	0,966667	0,06913	30	809
8	GCA_023315275.2_PDT001299634.2_genomic.fna	17	924	0	0,018066	1	0,035491	17	941
9	GCA_024452925.1_PDT001370221.1_genomic.fna	54	898	1	0,056723	0,981818	0,107249	55	952
10	GCA_024732165.1_PDT001378927.1_genomic.fna	40	818	4	0,04662	0,909091	0,088692	44	858
11	GCA_026006075.1_ASM2600607v1_genomic.fna	463	1021	32	0,311995	0,935354	0,467913	495	1484
12	GCA_026006095.1_ASM2600609v1_genomic.fna	500	992	33	0,335121	0,938086	0,493827	533	1492
13	GCA_026305215.1_PDT001493239.1_genomic.fna	1	541	0	0,001845	1	0,003683	1	542
14	GCA_031045075.2_PDT001744190.2_genomic.fna	50	905	3	0,052356	0,943396	0,099206	53	955
15	GCA_032567175.1_ASM3256717v1_genomic.fna	46	880	2	0,049676	0,958333	0,094456	48	926
16	GCA_033106465.1_PDT001960943.1_genomic.fna	51	915	5	0,052795	0,910714	0,099804	56	966
17	GCA_037052065.1_ASM3705206v1_genomic.fna	0	428	0	0	0	0	0	428
18	GCA_037203645.1_ASM3720364v1_genomic.fna	102	953	9	0,096682	0,918919	0,174957	111	1055
19	GCA_043678295.1_ASM4367829v1_genomic.fna	77	1048	3	0,068444	0,9625	0,127801	80	1125
20	GCA_043678335.1_ASM4367833v1_genomic.fna	274	1115	32	0,197264	0,895425	0,323304	306	1389
21	GCA_043678355.1_ASM4367835v1_genomic.fna	557	1020	46	0,353202	0,923715	0,511009	603	1577

22	GCA_043678375.1_ASM4367837v1_genomic.fna	442	1140	41	0,279393	0,915114	0,428087	483	1582
23	GCA_043950085.1_ASM4395008v1_genomic.fna	18	556	2	0,031359	0,9	0,060606	20	574
24	GCA_943323015.2_HRS-ES2-bin-440_genomic.fna	573	6318	3	0,083152	0,994792	0,153475	576	6891
25	GCA_943323415.2_HRS-ES4-bin-10_genomic.fna	2358	6664	8	0,261361	0,996619	0,41412	2366	9022
26	GCA_943323865.2_HRS-ES8-bin-118_genomic.fna	503	6897	1	0,067973	0,998016	0,127277	504	7400
27	GCA_943325355.2_HRSG-E10-bin-12_genomic.fna	8	826	0	0,009592	1	0,019002	8	834
28	GCA_943326675.2_HRSG-E12-bin-108_genomic.fna	68	781	3	0,080094	0,957746	0,147826	71	849
29	GCA_943327095.2_HRSG-E11-bin-71_genomic.fna	11	762	1	0,01423	0,916667	0,028025	12	773
30	GCA_943327695.2_HRS-HS3-bin-186_genomic.fna	1	111	0	0,008929	1	0,017699	1	112
31	GCA_943329525.2_HRS-HS6-bin-101_genomic.fna	107	3344	1	0,031006	0,990741	0,060129	108	3451
32	GCA_943329805.2_HRSG-H13-bin-193_genomic.fna	440	1702	2	0,205415	0,995475	0,340557	442	2142
33	GCA_943331045.1_HRSG-E2-bin-118_genomic.fna	16	145	0	0,099379	1	0,180791	16	161
34	GCA_943331235.2_HRSG-H12-bin-94_genomic.fna	24	539	0	0,042629	1	0,081772	24	563
35	GCA_943332185.1_HRS-HS9-bin-103_genomic.fna	66	1356	1	0,046414	0,985075	0,08865	67	1422
36	GCA_943332625.1_HRS-HS9-bin-149_genomic.fna	11	165	0	0,0625	1	0,117647	11	176
37	GCA_943333245.1_HRSG-E1-bin-8_genomic.fna	1671	10212	2	0,140621	0,998805	0,246533	1673	11883
38	GCA_943333305.2_HRSG-H6-bin-46_genomic.fna	25	2627	1	0,009427	0,961538	0,018671	26	2652
39	GCA_943334125.2_HRSG-E6-bin-23_genomic.fna	143	1636	1	0,080382	0,993056	0,148726	144	1779
40	GCA_943334475.2_HRSG-H3-bin-66_genomic.fna	188	4668	0	0,038715	1	0,074544	188	4856
41	GCA_943334715.2_HRSG-H6-bin-64_genomic.fna	55	163	1	0,252294	0,982143	0,40146	56	218
42	GCA_943334795.2_HRSG-H6-bin-28_genomic.fna	48	223	0	0,177122	1	0,30094	48	271
43	GCA_945860935.1_GE-03apr19-182_genomic.fna	0	379	0	0	0	0	0	379
44	GCA_945867275.1_TH-11nov19-195_genomic.fna	41	730	0	0,053178	1	0,100985	41	771

45	GCA_945870865.1_TrH-03may19-115_genomic.fna	108	1109	0	0,088743	1	0,163019	108	1217
46	GCA_945870955.1_MoE-23oct19-325_genomic.fna	66	768	0	0,079137	1	0,146667	66	834
47	GCA_945872895.1_MaE-04nov19-161_genomic.fna	16	837	0	0,018757	1	0,036824	16	853
48	GCA_945874295.1_AH-24oct19-10_genomic.fna	25	157	0	0,137363	1	0,241546	25	182
49	GCA_945877175.1_ZE-13nov19-95_genomic.fna	170	3432	4	0,047196	0,977011	0,090042	174	3602
50	GCA_945902215.1_MoH-23oct19-118_genomic.fna	203	1496	2	0,119482	0,990244	0,213235	205	1699

TS-CMCU con k= 31

idx	name	TP	FP	FN	precision	recall	f1	hh_exact	hh_est
1	GCA_006152045.1_ASM615204v1_genomic.fna	1	524	0	0,001905	1	0,003802	1	525
2	GCA_018421455.1_ASM1842145v1_genomic.fna	238	954	0	0,199664	1	0,332867	238	1192
3	GCA_020118255.1_ASM2011825v1_genomic.fna	56	1172	0	0,045603	1	0,087227	56	1228
4	GCA_021919605.1_PDT001092240.1_genomic.fna	42	916	0	0,043841	1	0,084	42	958
5	GCA_021953145.1_PDT001020446.1_genomic.fna	17	746	0	0,02228	1	0,04359	17	763
6	GCA_021972575.1_PDT001013406.1_genomic.fna	14	795	0	0,017305	1	0,034022	14	809
7	GCA_022062785.1_PDT000876120.1_genomic.fna	30	776	0	0,037221	1	0,07177	30	806
8	GCA_023315275.2_PDT001299634.2_genomic.fna	17	904	0	0,018458	1	0,036247	17	921
9	GCA_024452925.1_PDT001370221.1_genomic.fna	55	927	0	0,056008	1	0,106075	55	982
10	GCA_024732165.1_PDT001378927.1_genomic.fna	44	825	0	0,050633	1	0,096386	44	869
11	GCA_026006075.1_ASM2600607v1_genomic.fna	495	1004	0	0,33022	1	0,496489	495	1499
12	GCA_026006095.1_ASM2600609v1_genomic.fna	533	954	0	0,35844	1	0,527723	533	1487
13	GCA_026305215.1_PDT001493239.1_genomic.fna	1	555	0	0,001799	1	0,003591	1	556
14	GCA_031045075.2_PDT001744190.2_genomic.fna	53	910	0	0,055036	1	0,104331	53	963

15	GCA_032567175.1_ASM3256717v1_genomic.fna	48	855	0	0,053156	1	0,100946	48	903
16	GCA_033106465.1_PDT001960943.1_genomic.fna	56	940	0	0,056225	1	0,106464	56	996
17	GCA_037052065.1_ASM3705206v1_genomic.fna	0	411	0	0	0	0	0	411
18	GCA_037203645.1_ASM3720364v1_genomic.fna	111	943	0	0,105313	1	0,190558	111	1054
19	GCA_043678295.1_ASM4367829v1_genomic.fna	80	937	0	0,078663	1	0,145852	80	1017
20	GCA_043678335.1_ASM4367833v1_genomic.fna	306	980	0	0,237947	1	0,384422	306	1286
21	GCA_043678355.1_ASM4367835v1_genomic.fna	603	895	0	0,402537	1	0,574012	603	1498
22	GCA_043678375.1_ASM4367837v1_genomic.fna	483	1000	0	0,325691	1	0,491353	483	1483
23	GCA_043950085.1_ASM4395008v1_genomic.fna	20	555	0	0,034783	1	0,067227	20	575
24	GCA_943323015.2_HRS-ES2-bin-440_genomic.fna	576	3454	0	0,142928	1	0,250109	576	4030
25	GCA_943323415.2_HRS-ES4-bin-10_genomic.fna	2366	3622	0	0,395124	1	0,566435	2366	5988
26	GCA_943323865.2_HRS-ES8-bin-118_genomic.fna	504	3380	0	0,129763	1	0,229717	504	3884
27	GCA_943325355.2_HRSG-E10-bin-12_genomic.fna	8	656	0	0,012048	1	0,02381	8	664
28	GCA_943326675.2_HRSG-E12-bin-108_genomic.fna	71	664	0	0,096599	1	0,176179	71	735
29	GCA_943327095.2_HRSG-E11-bin-71_genomic.fna	12	403	0	0,028916	1	0,056206	12	415
30	GCA_943327695.2_HRS-HS3-bin-186_genomic.fna	1	106	0	0,009346	1	0,018519	1	107
31	GCA_943329525.2_HRS-HS6-bin-101_genomic.fna	108	1131	0	0,087167	1	0,160356	108	1239
32	GCA_943329805.2_HRSG-H13-bin-193_genomic.fna	442	1668	0	0,209479	1	0,346395	442	2110
33	GCA_943331045.1_HRSG-E2-bin-118_genomic.fna	16	116	0	0,121212	1	0,216216	16	132
34	GCA_943331235.2_HRSG-H12-bin-94_genomic.fna	24	496	0	0,046154	1	0,088235	24	520
35	GCA_943332185.1_HRS-HS9-bin-103_genomic.fna	67	1210	0	0,052467	1	0,099702	67	1277
36	GCA_943332625.1_HRS-HS9-bin-149_genomic.fna	11	157	0	0,065476	1	0,122905	11	168
37	GCA_943333245.1_HRSG-E1-bin-8_genomic.fna	1673	8581	0	0,163156	1	0,28054	1673	10254

38	GCA_943333305.2_HRSG-H6-bin-46_genomic.fna	26	2613	0	0,009852	1	0,019512	26	2639
39	GCA_943334125.2_HRSG-E6-bin-23_genomic.fna	144	1602	0	0,082474	1	0,152381	144	1746
40	GCA_943334475.2_HRSG-H3-bin-66_genomic.fna	188	4602	0	0,039248	1	0,075532	188	4790
41	GCA_943334715.2_HRSG-H6-bin-64_genomic.fna	56	145	0	0,278607	1	0,435798	56	201
42	GCA_943334795.2_HRSG-H6-bin-28_genomic.fna	48	210	0	0,186047	1	0,313725	48	258
43	GCA_945860935.1_GE-03apr19-182_genomic.fna	0	344	0	0	0	0	0	344
44	GCA_945867275.1_TH-11nov19-195_genomic.fna	41	699	0	0,055405	1	0,104994	41	740
45	GCA_945870865.1_TrH-03may19-115_genomic.fna	108	1014	0	0,096257	1	0,17561	108	1122
46	GCA_945870955.1_MoE-23oct19-325_genomic.fna	66	754	0	0,080488	1	0,148984	66	820
47	GCA_945872895.1_MaE-04nov19-161_genomic.fna	16	590	0	0,026403	1	0,051447	16	606
48	GCA_945874295.1_AH-24oct19-10_genomic.fna	25	150	0	0,142857	1	0,25	25	175
49	GCA_945877175.1_ZE-13nov19-95_genomic.fna	174	3410	0	0,048549	1	0,092602	174	3584
50	GCA_945902215.1_MoH-23oct19-118_genomic.fna	205	1472	0	0,122242	1	0,217853	205	1677

Count Sketch con k = 31

idx	**name**	**TP**	**FP**	**FN**	**precision**	**recall**	**f1**	**hh_exact**	**hh_est**
1	GCA_006152045.1_ASM615204v1_genomic.fna	1	515	0	0,001938	1	0,003868	1	516
2	GCA_018421455.1_ASM1842145v1_genomic.fna	231	973	7	0,19186	0,970588	0,320388	238	1204
3	GCA_020118255.1_ASM2011825v1_genomic.fna	53	1190	3	0,042639	0,946429	0,081601	56	1243
4	GCA_021919605.1_PDT001092240.1_genomic.fna	42	925	0	0,043433	1	0,083251	42	967
5	GCA_021953145.1_PDT001020446.1_genomic.fna	17	726	0	0,02288	1	0,044737	17	743
6	GCA_021972575.1_PDT001013406.1_genomic.fna	13	762	1	0,016774	0,928571	0,032953	14	775
7	GCA_022062785.1_PDT000876120.1_genomic.fna	29	780	1	0,035847	0,966667	0,06913	30	809

8	GCA_023315275.2_PDT001299634.2_genomic.fna	17	924	0	0,018066	1	0,035491	17	941
9	GCA_024452925.1_PDT001370221.1_genomic.fna	54	898	1	0,056723	0,981818	0,107249	55	952
10	GCA_024732165.1_PDT001378927.1_genomic.fna	40	818	4	0,04662	0,909091	0,088692	44	858
11	GCA_026006075.1_ASM2600607v1_genomic.fna	463	1021	32	0,311995	0,935354	0,467913	495	1484
12	GCA_026006095.1_ASM2600609v1_genomic.fna	500	992	33	0,335121	0,938086	0,493827	533	1492
13	GCA_026305215.1_PDT001493239.1_genomic.fna	1	541	0	0,001845	1	0,003683	1	542
14	GCA_031045075.2_PDT001744190.2_genomic.fna	50	905	3	0,052356	0,943396	0,099206	53	955
15	GCA_032567175.1_ASM3256717v1_genomic.fna	46	880	2	0,049676	0,958333	0,094456	48	926
16	GCA_033106465.1_PDT001960943.1_genomic.fna	51	915	5	0,052795	0,910714	0,099804	56	966
17	GCA_037052065.1_ASM3705206v1_genomic.fna	0	428	0	0	0	0	0	428
18	GCA_037203645.1_ASM3720364v1_genomic.fna	102	953	9	0,096682	0,918919	0,174957	111	1055
19	GCA_043678295.1_ASM4367829v1_genomic.fna	77	1048	3	0,068444	0,9625	0,127801	80	1125
20	GCA_043678335.1_ASM4367833v1_genomic.fna	274	1115	32	0,197264	0,895425	0,323304	306	1389
21	GCA_043678355.1_ASM4367835v1_genomic.fna	557	1020	46	0,353202	0,923715	0,511009	603	1577
22	GCA_043678375.1_ASM4367837v1_genomic.fna	442	1140	41	0,279393	0,915114	0,428087	483	1582
23	GCA_043950085.1_ASM4395008v1_genomic.fna	18	556	2	0,031359	0,9	0,060606	20	574
24	GCA_943323015.2_HRS-ES2-bin-440_genomic.fna	573	6318	3	0,083152	0,994792	0,153475	576	6891
25	GCA_943323415.2_HRS-ES4-bin-10_genomic.fna	2358	6664	8	0,261361	0,996619	0,41412	2366	9022
26	GCA_943323865.2_HRS-ES8-bin-118_genomic.fna	503	6897	1	0,067973	0,998016	0,127277	504	7400
27	GCA_943325355.2_HRSG-E10-bin-12_genomic.fna	8	826	0	0,009592	1	0,019002	8	834
28	GCA_943326675.2_HRSG-E12-bin-108_genomic.fna	68	781	3	0,080094	0,957746	0,147826	71	849
29	GCA_943327095.2_HRSG-E11-bin-71_genomic.fna	11	762	1	0,01423	0,916667	0,028025	12	773
30	GCA_943327695.2_HRS-HS3-bin-186_genomic.fna	1	111	0	0,008929	1	0,017699	1	112

31	GCA_943329525.2_HRS-HS6-bin-101_genomic.fna	107	3344	1	0,031006	0,990741	0,060129	108	3451
32	GCA_943329805.2_HRSG-H13-bin-193_genomic.fna	440	1702	2	0,205415	0,995475	0,340557	442	2142
33	GCA_943331045.1_HRSG-E2-bin-118_genomic.fna	16	145	0	0,099379	1	0,180791	16	161
34	GCA_943331235.2_HRSG-H12-bin-94_genomic.fna	24	539	0	0,042629	1	0,081772	24	563
35	GCA_943332185.1_HRS-HS9-bin-103_genomic.fna	66	1356	1	0,046414	0,985075	0,08865	67	1422
36	GCA_943332625.1_HRS-HS9-bin-149_genomic.fna	11	165	0	0,0625	1	0,117647	11	176
37	GCA_943333245.1_HRSG-E1-bin-8_genomic.fna	1671	10212	2	0,140621	0,998805	0,246533	1673	11883
38	GCA_943333305.2_HRSG-H6-bin-46_genomic.fna	25	2627	1	0,009427	0,961538	0,018671	26	2652
39	GCA_943334125.2_HRSG-E6-bin-23_genomic.fna	143	1636	1	0,080382	0,993056	0,148726	144	1779
40	GCA_943334475.2_HRSG-H3-bin-66_genomic.fna	188	4668	0	0,038715	1	0,074544	188	4856
41	GCA_943334715.2_HRSG-H6-bin-64_genomic.fna	55	163	1	0,252294	0,982143	0,40146	56	218
42	GCA_943334795.2_HRSG-H6-bin-28_genomic.fna	48	223	0	0,177122	1	0,30094	48	271
43	GCA_945860935.1_GE-03apr19-182_genomic.fna	0	379	0	0	0	0	0	379
44	GCA_945867275.1_TH-11nov19-195_genomic.fna	41	730	0	0,053178	1	0,100985	41	771
45	GCA_945870865.1_TrH-03may19-115_genomic.fna	108	1109	0	0,088743	1	0,163019	108	1217
46	GCA_945870955.1_MoE-23oct19-325_genomic.fna	66	768	0	0,079137	1	0,146667	66	834
47	GCA_945872895.1_MaE-04nov19-161_genomic.fna	16	837	0	0,018757	1	0,036824	16	853
48	GCA_945874295.1_AH-24oct19-10_genomic.fna	25	157	0	0,137363	1	0,241546	25	182
49	GCA_945877175.1_ZE-13nov19-95_genomic.fna	170	3432	4	0,047196	0,977011	0,090042	174	3602
50	GCA_945902215.1_MoH-23oct19-118_genomic.fna	203	1496	2	0,119482	0,990244	0,213235	205	1699

Gráfico 1: Comparación entre los heavy hitters de k-mers 21 con phi 2e-06

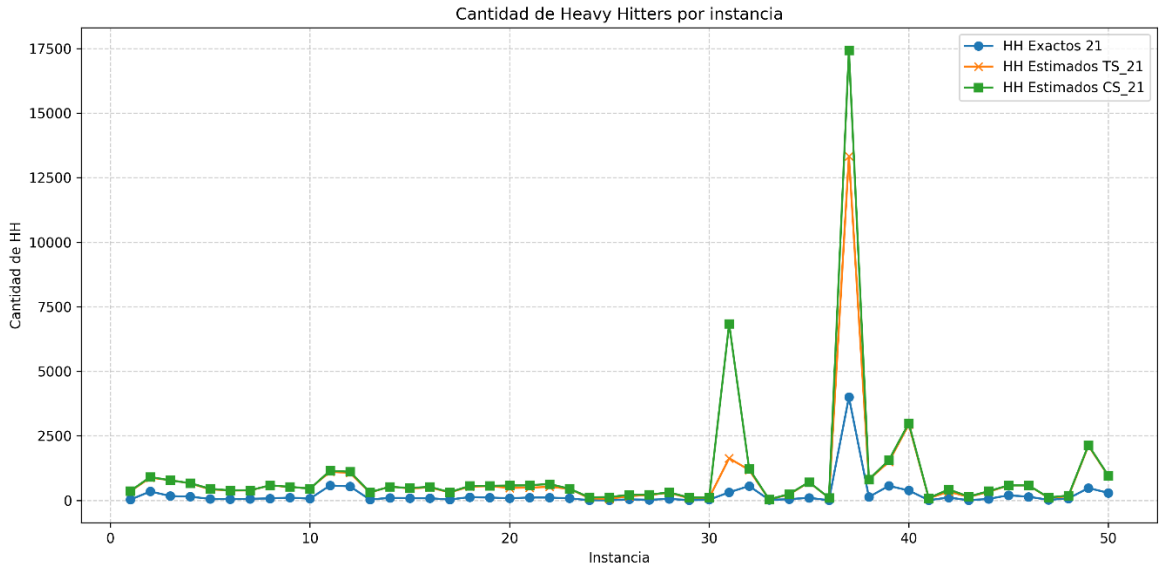
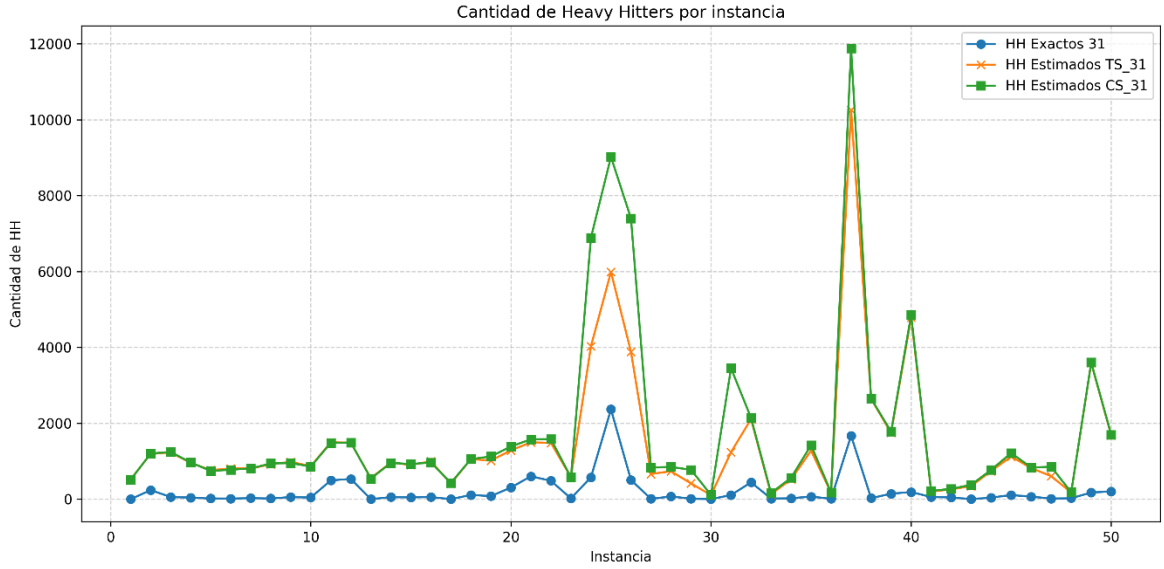


Gráfico 2: Comparación entre los heavy hitters de k-mers 31 con phi 2e-06



Análisis comparativo

Estructura y funcionamiento

- CountSketch (CS):
    - Usa funciones hash y un conjunto de contadores con signo (+1, -1).
    - La estimación se obtiene tomando la mediana de los valores acumulados.
    - Permite compensar ruido y manejar bien distribuciones sesgadas, reduciendo el sesgo de las colisiones.
- TowerSketch (TS-CMCU):
    - Extiende CountMin Sketch, organizando los contadores en “niveles” o torres de diferente tamaño.
    - Aplica conservative update, es decir, al insertar un k-mer solo se actualizan los contadores mínimos, evitando inflar artificialmente las frecuencias.
    - Maneja mejor los heavy hitters y distribuye el error de manera controlada.

Precisión y error

- CS: ofrece estimaciones simétricas, lo que reduce el error relativo, pero puede generar más variabilidad en los resultados.
- TS-CMCU: tiende a mejorar la precisión práctica gracias al conservative update, reduciendo falsos positivos en los heavy hitters.

Uso de memoria

- CS: Requiere un tamaño de sketch definido por parámetros (d, w). El error decrece con más memoria, pero no diferencia la importancia entre elementos.
- TS-CMCU: Usa memoria de forma más eficiente porque concentra contadores en distintos niveles, logrando un mejor balance entre exactitud y espacio.

Ventajas y desventajas

- CS:
    - Buen rendimiento teórico para distribuciones con gran dispersión.
    - Simétrico: maneja bien valores positivos y negativos.

    - Puede tener alta varianza en la estimación.
- TS-CMCU:
    - Más preciso en la práctica para heavy hitters.
    - Conservative update evita sobreestimación.
    - Mayor complejidad de implementación y ajuste de parámetros.

Aplicación en k-mers

- CS: adecuado cuando se requiere una primera aproximación rápida al conteo de k-mers frecuentes en secuencias genómicas.
- TS-CMCU: preferible cuando se busca un equilibrio más fuerte entre bajo error y uso razonable de memoria, especialmente útil en conjuntos de datos grandes y ruidosos.

Conclusión

Los resultados obtenidos muestran que el algoritmo logra identificar un número considerable de heavy hitters en los archivos genómicos, aunque con un nivel variable de exactitud. Se observa que, en la mayoría de los casos, el recall es cercano a 1, lo que indica que casi todos los heavy hitters reales son detectados. Sin embargo, la precisión es significativamente menor, reflejando que junto a los verdaderos positivos aparecen también muchos falsos positivos. Esto provoca que el valor F1, que equilibra precisión y recall, sea generalmente bajo.

En términos prácticos, esto significa que el método es eficaz para no perder k-mers frecuentes (alta sensibilidad), pero menos fiable para discriminar cuáles son realmente representativos (baja especificidad). Por lo tanto, se requiere una calibración más fina de los parámetros del sketch y posiblemente la combinación con otras técnicas, a fin de reducir falsos positivos y mejorar la precisión sin sacrificar el recall.